

PLEASE TYPE**THE UNIVERSITY OF NEW SOUTH WALES**
Thesis/Dissertation SheetSurname or Family name: **Yau**First name: **Sheree**

Other name/s:

Abbreviation for degree as given in the University calendar: **PhD**School: **Biotechnology and Biomolecular Sciences**Faculty: **Faculty of Science**Title: **Molecular microbial ecology of Antarctic lakes****Abstract 350 words maximum: (PLEASE TYPE)**

The Vestfold Hills is a coastal Antarctic oasis, a rare ice-free region on the continent containing hundreds of marine-derived lakes. These lakes are microbially-dominated systems constrained by extremes of cold, salinity and light availability. Most Antarctic lakes are ice-covered for the majority of the year and are thus largely closed systems that often become meromictic (permanently stratified). The physical and chemical gradients that exist within an isolated system makes it possible to relate microbial taxa to abiotic variables. These factors make Antarctic lakes ideal model ecosystems to study microbial diversity, evolution and influence on geochemistry.

Sequencing of ribosomal genes from the environment has revolutionised microbial ecology by revealing the immense diversity of microbial life. However, this approach does not directly describe the physiology and ecological roles of members in a community. Random sequencing of genetic material from the environment (metagenomics) allows the determination not only of the microbial composition, but also its metabolic potential. Historic, physical, chemical and biological data available for two meromictic lakes in the Vestfold Hills, Ace Lake and Organic Lake, indicate each has unique microbial populations and biogeochemical properties. Metagenomic sequencing was applied to these two lakes to gain insight into their microbial ecology. Analytical methods to support metagenomic inferences were also developed and applied. These included identification and quantification of proteins from environmental samples (metaproteomics), which indicates active community members and biochemical processes; as well as microscopy, for determination microbial/viral abundances and morphology.

Analysis of these lake ecosystems yielded extensive genetic information from taxa previously unknown in the lakes, in particular, phycodnaviruses and a member of the newly described virophage viral family. The combination of metagenomics, metaproteomics and physico-chemical data enabled the ecological roles of taxa in the lakes and potential mechanisms of adaptation to the Antarctic environment to be determined. These analyses revealed viral predation, higher propensity of nutrient recycling and strategies of carbon conservation in Antarctic lake ecosystems. These molecular-based discoveries allowed the role of previously unrecognised taxa and metabolic processes to be modelled enabling implication to be drawn about Antarctic and other aquatic environments.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

.....
Signature.....
Witness.....
Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY

Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

Molecular microbial ecology of Antarctic lakes

Sheree Yau

A thesis in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Biotechnology and Biomolecular Sciences
Faculty of Science
University of New South Wales, Australia

February, 2013

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed

Date

Abstract

The Vestfold Hills is a coastal Antarctic oasis, a rare ice-free region on the continent containing hundreds of marine-derived lakes. These lakes are microbially-dominated systems constrained by extremes of cold, salinity and light availability. Most Antarctic lakes are ice-covered for the majority of the year and are thus largely closed systems that often become meromictic (permanently stratified). The physical and chemical gradients that exist within an isolated system makes it possible to relate microbial taxa to abiotic variables. These factors make Antarctic lakes ideal model ecosystems to study microbial diversity, evolution and influence on geochemistry.

Sequencing of ribosomal genes from the environment has revolutionised microbial ecology by revealing the immense diversity of microbial life. However, this approach does not directly describe the physiology and ecological roles of members in a community. Random sequencing of genetic material from the environment (metagenomics) allows the determination not only of the microbial composition, but also its metabolic potential. Historic, physical, chemical and biological data available for two meromictic lakes in the Vestfold Hills, Ace Lake and Organic Lake, indicate each has unique microbial populations and biogeochemical properties. Metagenomic sequencing was applied to these two lakes to gain insight into their microbial ecology. Analytical methods to support metagenomic inferences were also developed and applied. These included identification and quantification of proteins from environmental samples (metaproteomics), which indicates active community members and biochemical processes; as well as microscopy, for determination microbial/viral abundances and morphology.

Analysis of these lake ecosystems yielded extensive genetic information from taxa previously unknown in the lakes, in particular, phycodnaviruses and a member of the newly described virophage viral family. The combination of metagenomics, metaproteomics and physico-chemical data enabled the ecological roles of taxa in the lakes and potential mechanisms of adaptation to the Antarctic environment to be determined. These analyses revealed viral predation, higher propensity of nutrient recycling and strategies of carbon conservation in Antarctic lake ecosystems. These molecular-based discoveries allowed the role of previously unrecognised taxa and metabolic processes to be modelled enabling implication to be drawn about Antarctic and other aquatic environments.

Acknowledgements

List of Publications

In publications arising from my PhD work, my supervisor Prof Ricardo Cavicchioli and my co-supervisor Dr Federico Lauro were involved in the research design and editing of the manuscripts. Where versions of published material, or material submitted for publication appears in this thesis, details of the contributions made by myself and others precede it.

- **Sheree Yau**, Federico M. Lauro, Timothy J. Williams, Matthew Z. DeMaere, Mark V. Brown, John Rich, John A.E. Gibson, Ricardo Cavicchioli. Strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline lake. *The ISME Journal* (submitted), 2013.
- Khawar S. Siddiqui, Timothy J. Williams, David Wilkins, **Sheree Yau**, Michelle A. Allen, Mark V. Brown, Federico M. Lauro, Ricardo Cavicchioli. Psychrophiles. *Annual Review of Earth and Planetary Sciences* (doi: 10.1146/annurev-earth-040610-133514), 2013.
- David Wilkins, **Sheree Yau**, Timothy J. Williams, Michelle Allen, Mark V. Brown, Matthew Z. DeMaere, Federico M. Lauro and Ricardo Cavicchioli. Key Microbial Drivers in Antarctic Aquatic Environments. *FEMS Microbiology Reviews* (doi:10.1111/1574-6976.12007), 2012.
- **Sheree Yau** and Ricardo Cavicchioli. Microbial communities in Antarctic lakes: Entirely new perspectives from metagenomics and metaproteomics. *Microbiology Australia* 32:157–159, 2011.
- **Sheree Yau**, Federico M. Lauro, Matthew Z. DeMaere, Mark V. Brown, Torsten Thomas, Mark J. Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffrey M. Hoffman, John A. Gibson and Ricardo Cavicchioli. Virophage control of antarctic algal host–virus dynamics. *Proceedings of the National Academy of Sciences USA* 108:6163–6168, 2011.
- Federico M. Lauro, Matthew Z. DeMaere, **Sheree Yau**, Mark V. Brown, Charmaine Ng, David Wilkins, Mark J. Raftery, John A.E. Gibson, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffrey M. Hoffman, Torsten Thomas and Ricardo Cavicchioli. An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME Journal* 5:879–895, 2011.

Contents

| | |
|---|-----------|
| 1 General introduction | 1 |
| Co-authorship statement | 1 |
| 1.1 Antarctic lakes | 2 |
| 1.2 The Vestfold Hills | 3 |
| 1.3 Insights from molecular studies of Antarctic lakes | 3 |
| 1.3.1 Bacterial diversity: adaptation to unique physical and chemical conditions | 5 |
| 1.3.2 <i>Archaea</i> : methanogens and haloarchaea | 7 |
| 1.3.3 <i>Eucarya</i> perform multiple ecosystem roles | 8 |
| 1.4 Integrative studies to derive whole ecosystem function | 9 |
| 1.4.1 A single gene approach | 9 |
| 1.4.2 ‘-omics’ approaches | 10 |
| 1.5 Objectives | 12 |
| 2 Development of methods to complement metagenomic sequencing for an integrative study of Ace Lake | 13 |
| Co-authorship statement | 13 |
| 2.1 Abstract | 14 |
| 2.2 Introduction | 14 |
| 2.3 Materials and methods | 16 |
| 2.3.1 Ace Lake samples | 16 |
| 2.3.2 DNA extraction, sequencing and data cleanup | 17 |
| 2.3.3 Metagenomic DNA assembly and annotation | 18 |
| 2.3.4 Epifluorescence microscopy | 18 |
| 2.3.5 Protein extraction | 18 |
| 2.3.6 1D-SDS PAGE and LC-MS-MS | 19 |
| 2.3.7 Metaproteomic mass spectra analysis | 19 |
| 2.4 Results and discussion | 20 |
| 2.4.1 Development of an epifluorescence microscopy method | 20 |
| 2.4.2 Community stratification supported by cell and VLP densities . . | 22 |
| 2.4.3 Development of a metaproteomic mass spectra analysis workflow | 24 |
| 2.4.4 Insights from the metaproteomic analysis of Ace Lake | 29 |
| 2.5 Conclusions | 32 |

| | |
|--|-----------|
| 3 Virophage control of Antarctic algal host–virus dynamics | 33 |
| Co-authorship statement | 33 |
| 3.1 Abstract | 34 |
| 3.2 Introduction | 35 |
| 3.3 Materials and methods | 36 |
| 3.3.1 Samples and DNA sequencing | 36 |
| 3.3.2 Transmission electron microscopy | 36 |
| 3.3.3 Metagenomic assembly and annotation | 36 |
| 3.3.4 Genome completion and annotation | 37 |
| 3.3.5 Phylogenetic analysis | 38 |
| 3.3.6 Metaproteomic analysis | 38 |
| 3.3.7 Model of algal host–virus and virophage dynamics | 38 |
| 3.4 Results and discussion | 39 |
| 3.4.1 Dominance of phycodnaviruses in Organic Lake | 39 |
| 3.4.2 Complete genome of an Organic Lake virophage | 44 |
| 3.4.3 Gene exchange between virophage and phycodnaviruses | 51 |
| 3.4.4 Virophage in algal host–phycodnavirus dynamics | 52 |
| 3.4.5 Ecological relevance of virophages in aquatic systems | 54 |
| 3.5 Acknowledgements | 57 |
| 4 Strategies of carbon conservation and unusual sulphur biogeochemistry in Organic Lake | 59 |
| Co-authorship statement | 59 |
| 4.1 Abstract | 60 |
| 4.2 Introduction | 61 |
| 4.3 Materials and methods | 62 |
| 4.3.1 Characteristics of the lake and sample collection | 62 |
| 4.3.2 Physical and chemical analyses | 63 |
| 4.3.3 Epifluorescence microscopy | 63 |
| 4.3.4 Cellular diversity analyses | 64 |
| 4.3.5 Analysis of functional potential | 64 |
| 4.4 Results and discussion | 68 |
| 4.4.1 Abiotic properties and water column structure | 68 |
| 4.4.2 Overall microbial diversity | 70 |
| 4.4.3 Variation of microbial composition according to size and depth . | 72 |
| 4.4.4 Organic Lake functional potential | 78 |
| 4.4.5 Carbon resourcefulness in dominant heterotrophic bacteria . . . | 79 |
| 4.4.6 Regenerated nitrogen is predominant in the nitrogen cycle . . . | 86 |
| 4.4.7 Molecular basis for unusual sulphur chemistry | 89 |
| 4.5 Conclusions | 97 |

| | |
|---|------------|
| 5 General discussion, future work and conclusions | 99 |
| 5.1 Possibile future work on Organic Lake | 99 |
| 5.1.1 Organic Lake community dynamics | 100 |
| 5.1.2 OLV physiology and ecology | 102 |
| 5.1.3 Organic lake biogeochemistry | 104 |
| 5.2 Perspectives on ‘-omics’ approaches | 106 |
| 5.2.1 Next generation sequencing technologies | 106 |
| 5.2.2 Emerging bioinformatic bottlenecks | 107 |
| 5.2.3 Prospects for closing the bioinformatic gap | 108 |
| 5.2.4 ‘omes are only as good as our databases | 111 |
| 5.2.5 Metagenomes and metaproteomes are time capsules | 113 |
| 5.3 Concluding remarks | 113 |
| References | 135 |
| Appendix A PCR-based studies of Antarctic Lakes | 137 |
| Appendix B Proteins identified in Ace Lake metaproteomic analysis | 147 |
| Appendix C Peptide sequences of OLV and OLPV proteins identified in the metaproteome | 213 |
| Appendix D Microbial taxa detected in the Organic Lake water column | 215 |

List of Figures

| | | |
|------|---|----|
| 1.1 | Map of the Vestfold Hills | 4 |
| 2.1 | Physico-chemical and biological structure of Ace Lake from (Lauro <i>et al.</i> , 2011) | 15 |
| 2.2 | Epifluorescence microscopy of Ace Lake microbiota | 23 |
| 2.3 | Counts of microbial cells and VLPs in Ace Lake | 24 |
| 2.4 | Lack of VLP in 5 m and 12.7 m Ace Lake samples | 25 |
| 2.5 | General shotgun proteomics workflow | 26 |
| 2.6 | Statistical analysis of Ace Lake metaproteome | 29 |
| 3.1 | Plot of percent GC content <i>vs</i> coverage for Organic Lake scaffolds | 40 |
| 3.2 | Transmission electron micrographs of Organic Lake VLPs | 41 |
| 3.3 | Phylogeny of OLPV B family DNA polymerase sequences | 42 |
| 3.4 | Phylogeny of the 18S rRNA genes from Organic Lake | 43 |
| 3.5 | Maps of OLPV genomic scaffolds | 44 |
| 3.6 | Phylogeny of OLPV major capsid protein sequences | 45 |
| 3.7 | Genomic map of Organic Lake virophage | 47 |
| 3.8 | Genomic comparison of OLV and OLPVs | 52 |
| 3.9 | Lotka-Volterra models of host–OLPV–OLV population dynamics | 53 |
| 3.10 | Virophage capsid proteins in environmental samples | 54 |
| 3.11 | Phylogeny of virophage capsid proteins | 56 |
| 4.1 | Vertical profiles of <i>in situ</i> Organic Lake abiotic parameters | 68 |
| 4.2 | Bathymetry of Organic Lake | 69 |
| 4.3 | Vertical structure of Organic Lake | 69 |
| 4.4 | Epifluorescence microscopy images of Organic Lake microbiota | 71 |
| 4.5 | PCA of physico-chemical parameters | 72 |
| 4.6 | Diversity of <i>Bacteria</i> in Organic Lake | 73 |
| 4.7 | Diversity of <i>Eucarya</i> in Organic Lake | 74 |
| 4.8 | Heatmap of Organic Lake SSU composition | 76 |
| 4.9 | Vertical profiles of potential for carbon cycling in Organic Lake | 79 |
| 4.10 | Phylogeny of rhodopsin homologues | 83 |
| 4.11 | Maps of OL-R1 rhodopsin-containing scaffolds | 84 |
| 4.12 | Vertical profiles of potential for nitrogen cycling in Organic Lake | 86 |
| 4.13 | Vertical profiles of potential for sulphur cycling in Organic Lake | 89 |

| | | |
|------|--|-----|
| 4.14 | Phylogeny of DddD DMSP lyase homologues | 92 |
| 4.15 | Maps of OL-dddD-containing scaffolds | 93 |
| 4.16 | Phylogeny of DddL DMSP lyase homologues | 94 |
| 4.17 | Phylogeny of DddP DMSP lyase homologues | 95 |
| 4.18 | Phylogeny of DmdA DMSP demethylase homologues | 96 |
| 5.1 | Plot of sequencing cost, curated proteins and metagenomic publications from Temperton and Giovannoni (2012) | 112 |

List of Tables

| | | |
|-----|--|-----|
| 2.1 | Summary of metagenomic data for Ace Lake profile | 17 |
| 2.2 | Comparison of peptides/proteins identified with NR <i>vs.</i> matched metagenomes | 28 |
| 3.1 | List of primers used to close the OLV genome. | 37 |
| 3.2 | Summary of metagenomic data for Organic Lake 0.1 µm samples | 40 |
| 3.3 | OLPV and OLV proteins identified in the metaproteome | 46 |
| 3.4 | Annotation of Organic Lake virophage genome | 48 |
| 3.5 | OLV MCP in Organic Lake and Ace Lake and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) | 55 |
| 4.1 | Summary of metagenomic data for Organic Lake profile | 62 |
| 4.2 | List of KO groups searched for in the Organic lake metagenome | 65 |
| 4.3 | List of query functional marker gene sequences | 67 |
| 4.4 | Physico-chemical properties of Organic Lake profile | 70 |
| 4.5 | Taxonomic origin of genes involved carbon cycling | 80 |
| 4.6 | Counts of genes involved in dimethylsulphopropionate (DMSP) catabolism and photoheterotrophy in aquatic metagenomes | 84 |
| 4.7 | Taxonomic origin of genes involved in nitrogen conversions | 88 |
| 4.8 | Taxonomic origin of genes involved in sulphur conversions | 91 |
| 5.1 | Comparison of next generation DNA sequencing platforms | 107 |
| A.1 | PCR-based studies of Antarctic Lakes | 138 |
| B.1 | Proteins identitfied in the Ace Lake 5 m sample 0.1 µm size-fraction metaproteome | 148 |
| B.2 | Proteins identitfied in the Ace Lake 11.5 m sample 0.1 µm size-fraction metaproteome | 165 |
| B.3 | Proteins identitfied in the Ace Lake 12.7 m sample 0.1 µm size-fraction metaproteome | 173 |
| B.4 | Proteins identitfied in the Ace Lake 14 m sample 0.1 µm size-fraction metaproteome | 191 |
| B.5 | Proteins identitfied in the Ace Lake 18 m sample 0.1 µm size-fraction metaproteome | 204 |
| B.6 | Proteins identitfied in the Ace Lake 23 m sample 0.1 µm size-fraction metaproteome | 208 |

| | |
|---|-----|
| C.1 Peptide data for Organic Lake metaproteomics analysis | 214 |
| D.1 Microbial taxa detected in the Organic Lake profile | 215 |

List of Abbreviations

1D-SDS PAGE one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis

AAnP aerobic anoxygenic photosynthesis

ABC ATP-binding cassette

ACT Artemis comparison tool

ALV Ace Lake virophage

EC2 Elastic Compute Cloud

ANOSIM analysis of similarity

ApMV *Acanthamoeba polyphaga* mimivirus

BchlA bacteriochlorophyll A

BLAST basic local alignment search tool

CAMERA Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis

CAS CRISPR-associated proteins

COG clusters of orthologous groups

CPU central processing unit

CRISPR clustered regularly interspaced short palindromic repeat

CroV *Cafeteria roenbergensis* virus

DGGE denaturing gradient gel electrophoresis

DMS dimethylsulphide

DMSO dimethylsulphoxide

DMSP dimethylsulphopropionate

DPOB DNA polymerase B

DO dissolved oxygen
DOC dissolved organic carbon
DRP dissolved reactive phosphorus
FDR false discovery rate
GOS global ocean sampling
GPU graphical processing unit
GSB green sulphur bacteria
HMMER biosequence analysis using profile hidden Markov models
HPLC high performance liquid chromatography
IMG/M Integrated Microbial Genomes and Metagenomes
JCVI J. Craig Venter Institute
KEGG Kyoto Encyclopedia of Genes and Genomes
KO KEGG Orthology
KOBAS KEGG Orthology Based Annotation System
LC liquid chromatography
MCP major capsid protein
MEGA Molecular Evolutionary Genetic Analysis
MG-RAST Metagenome Rapid Annotation using Subsystems Technology
MS mass spectrometry
MS-MS two dimensional mass spectrometry
NCBI National Center for Biotechnology Information
NCLDV nucleo-cytoplasmic large DNA virus
NGS next generation sequencing
NR non-redundant database
NSA normalised spectral abundance
OLPV Organic Lake phycodnavirus
OLV Organic Lake virophage
ORF open-reading frame

OTU operational taxonomic unit

PCA principal component analysis

PCR polymerase chain reaction

PCTE polycarbonate Track Etch

PR proteorhodopsin

PV phycodnavirus

QIIME Quanitative Insights Into Microbial Ecology

RDP Ribosomal Database Project

RE restriction endonuclease

RM restriction modification

rRNA ribosomal RNA

rTCA reverse tricarboxylic acid

RuBisCO ribulose-bisphosphate carboxylase oxygenase

SAG single-cell amplified genome

SIP stable isotope probing

SO Southern Ocean

SRB sulphate-reducing bacteria

SSU small subunit ribosomal RNA

STAMP Statistical Analysis of Metagenomic Profiles

SVG single virus genome

TEM transmission electron microscopy

TDN total dissolved nitrogen

TDP total dissolved phosphorus

TDS total dissolved sulphur

TIGRFAM the Institute of Genomic Research curated protein database

TN total nitrogen

TOC total organic carbon

TP total phosphorus

TS total sulphur

VLP virus-like particle

VIROME Viral Informatics Resource for Metagenome Exploration

WGS whole genome shotgun

WL Wood-Ljungdahl; or reductive acetyl-CoA

Chapter 1

General introduction

Co-authorship statement

Sections of this chapter have been published as:

David Wilkins, **Sheree Yau**, Timothy Williams, Michelle Allen, Mark V. Brown, Matthew Z. DeMaere, Federico M. Lauro and Ricardo Cavicchioli. Key Microbial Drivers in Antarctic Aquatic Environments. *FEMS Microbiology Reviews* (doi: 10.1111/1574-6976.12007), 2012.

I contributed the section of the publication entitled *Antarctic lakes* excluding the subsection, *Microbial mats as microcosms of Antarctic life*. This material appears in sections of this introduction.

Antarctica is a “frozen desert” of constant low temperature, little precipitation and is subject to the polar light cycle where only specially adapted organisms can survive. The continent is covered by ice up to 4 km thick that spans 13.8 million km². A tiny 0.32% of the land area is ice-free, most of which consists of exposed rocky peaks or nunataks such as in the Ellsworth, the Transantarctic and the North Victoria Land Mountains. Only 1–2% of that ice-free land is found in coastal oases; however, it is these regions where Antarctic life is concentrated (Hodgson, 2012). They are breeding sites for large animals such as seals, penguins and sea birds and some of the only locations where plants and lichens are found. Coastal oases are also distinguished by the presence of hundreds of lakes and ponds. Life in these lakes is microbially dominated with few, if any, metazoan inhabitants (Laybourn-Parry, 1997) making them ideal locations to study Antarctic microbiota. The lakes span a continuum of environmental factors such as salinity and are “natural laboratories” to examine adaptations to a property of interest. The reduced biodiversity of Antarctic lakes makes them ideal model systems to examine microbial influence on geochemistry as it is possible to encompass a large proportion of the diversity present using molecular methods and relate taxa to particular processes (Laybourn-Parry and Pearce, 2007).

This introduction will review molecular-based microbiological research on Antarctic lakes. As this thesis focused on planktonic communities from two lakes in the Vestfold Hills, emphasis will be given to describing research from these study sites and to microbial ecology of the water column.

1.1 Antarctic lakes

In Antarctica, perenially available liquid water is found predominantly in lakes. These range from subglacial lakes that are sealed beneath kilometres of ice to completely exposed rock-bound lakes. The majority of lakes are found in the coastal oases. In East Antarctica these include the Vestfold Hills, Bungar Hills, Larsemann Hills, Syowa Oasis, Schirmacher Oasis, Grearson Hills and McMurdo Dry Valleys. In West Antarctic, the Antarctic Peninsula, the sub-Antarctic islands and maritime islands house multiple lakes. Of these locations, the best studied lake systems are those of the McMurdo Dry Valleys, The Vestfold Hills and the sub-Antarctic islands.

These lakes span a wide range of physical and chemical properties from freshwater to hypersaline and constantly ice-covered to melted. Some are permanently stratified and termed meromictic if they thaw seasonally, or amictic if they are always ice-covered. Stratified lakes provide a unique opportunity to describe aquatic microbial populations along steep chemical gradients and taxa can be related to the properties of that layer. They are also of particular interest because the anoxic bottom waters help preserve a paleogeological record in the sediments of geological and climatic changes. Most lakes are ice-covered for the majority of the year making them effectively isolated, and some may be truly closed systems if ice-cover is permanent. The age of water varies considerably; for example, outflow of subglacial water at Blood Falls is estimated to be 1.5 million years old (Mikucki *et al.*, 2009) while water from Lake Miers is less than

300 years old (Green *et al.*, 1988).

Most of these lakes were formed when the retreat of the continental ice-shelf lead to isostatic uplift of the land (Burton, 1981). As a result, the majority of lakes in the coastal oases are composed of relic seawater and are predominantly saline or hypersaline (Burke and Burton, 1988). In the latter, salinity is high due to concentrated by ablation (evaporation and sublimination). Lakes close to the coastline, such as Lake Rookery in the Vestfold Hills, may still occasionally experience marine inputs (Burton, 1981).

Freshwater lakes near the continental ice shelf were likely already above sea-level as the ice receded and are not of marine origin (Bronge, 2004). Other freshwater lakes were originally marine-derived but have been flushed fresh by glacial meltwater (Pickard *et al.*, 1986). The chemistry of the exposed lakes is very much influenced by the water balance from local geographic and climatic conditions. Input sources include precipitation from the ice-shelf and glacial melt streams (Burton, 1981).

1.2 The Vestfold Hills

The Vestfold Hills (Figure 1.1) is a ice-free region of approximately 400 km² on the eastern shore of the Prydz Bay, East Antarctica in the Australian Antarctic Territory (Gibson, 1999). The region was first sighted and named in 1935 (Law, 1959). Only intermittent expeditions occurred in the area until the establishment of Davis Station (68°33'S, 78°15'E) in 1957 (Law, 1959). The Vestfold Hills are made up of three large peninsulae, Broad, Mule and Long Peninsula, separated by fjords connected to the sea. Some of these fjords are large, such as Ellis Fjord, which is 10 km long, up to 100 m deep and has become a stratified system due to its restricted opening to the ocean (Burke and Burton, 1988). The region was formed approximately 10,000 years ago in the early Holocene as the continental ice receded and the rocky peninsulae rose above sea-level (Zwartz *et al.*, 1998).

When first discovered, the Vestfold Hills was immediately noted for its extensive ice-free land and the numerous lakes (Johnstone *et al.*, 1973). The Australian Antarctic Data Centre lists more than 3,000 water bodies mapped in the Vestfold Hills, ranging in area from 1 to 8,757,944 m². More than 300 lakes and ponds have been described, including approximately 20% of the world's meromictic lakes (Gibson, 1999).

1.3 Insights from molecular studies of Antarctic lakes

The majority of molecular-based studies of Antarctic lake microbial communities have made use of polymerase chain reaction (PCR) amplification of small subunit ribosomal RNA (SSU) sequences to survey the diversity of *Bacteria* and in some cases *Archaea* and *Eucarya*. A comprehensive list of PCR-based studies of Antarctic lakes and their major findings are provided in Appendix A. Microbial composition has been determined by cloning and sequencing of ribosomal RNA (rRNA) gene amplicons (Bowman *et al.*, 2000b,a; Gordon *et al.*, 2000; Christner *et al.*, 2001; Purdy *et al.*, 2003; Karr *et al.*, 2006; Matsuzaki *et al.*, 2006; Kurosawa *et al.*, 2010; Bielewicz *et al.*, 2011), although



Figure 1.1: A map of the Vestfold Hills showing fjords, bays and lakes (numbered). The Southern Ocean is shown in grey, meromictic lakes coloured in black, seasonally isolated lakes and basins are striped and the continental ice-shelf is stippled. Inset is the position of the Vestfold Hills relative to Australia and to the Antarctic coastal oases. The lakes are: (1) unnamed lake 2, (2) Organic Lake, (3) Pendant Lake, (4) Glider Lake, (5) Ace Lake, (6) unnamed lake 1, (7) Williams Lake, (8) Abraxas Lake, (9) Johnstone Lake, (10) Ekho Lake, (11) Lake Farrell, (12) Shield Lake, (13) Oval Lake, (14) Ephyra Lake, (15) Scale Lake, (16) Lake Anderson, (17) Oblong Lake, (18) Lake McCallum, (19) Clear Lake, (20) Laternula Lake and (21) South Angle Lake. Map image is from Gibson (1999) with some minor modifications.

many studies have also made use of denaturing gradient gel electrophoresis (DGGE) to provide a molecular “fingerprint” of the community (Pearce, 2003; Pearce *et al.*, 2003; Karr *et al.*, 2005; Pearce, 2005; Pearce *et al.*, 2005; Unrein *et al.*, 2005; Glatz *et al.*, 2006; Mikucki and Priscu, 2007; Mosier *et al.*, 2007; Schiaffino *et al.*, 2009; Villaescusa *et al.*, 2010). Functional genes have also been targeted using PCR amplification to assess the potential of biochemical processes occurring, such as nitrogen fixation (Olson *et al.*, 1998), ammonia oxidation (Voytek *et al.*, 1999), anoxygenic photosynthesis (Karr *et al.*, 2003), and dissimilatory sulfite reduction (Karr *et al.*, 2005; Mikucki *et al.*, 2009). These PCR-based analyses of Antarctic lake communities have shed light on microbial diversity, temporal and spatial distributions and started to delve into the key drivers in whole ecosystem function.

Relatively few metagenomic studies, which entail random high-throughput sequencing of environmental DNA, have been published on Antarctic lakes (López-Bueno *et al.*, 2009; Ng *et al.*, 2010; Lauro *et al.*, 2011; Yau *et al.*, 2011). However, these few studies have been able to assess both the taxonomic composition and genetic potential of lake communities, and in many cases have linked function to specific members of the community. When coupled with functional “-omic” techniques (to date metaproteomics has been applied, but not metatranscriptomics or stable isotope probing), information has also been gained about the genetic complement expressed by the resident populations (Ng *et al.*, 2010; Lauro *et al.*, 2011; Yau *et al.*, 2011). These studies are described in section 1.4 of this introduction and in the body of this thesis.

1.3.1 Bacterial diversity: adaptation to unique physical and chemical conditions

Most molecular studies conducted on Antarctic lakes have focused on *Bacteria*. Consistent with the wide range of physical and chemical properties of Antarctic lakes, a large variation in species assemblages have been found. While exchange of microorganisms must be able to occur between lakes that are in close vicinity to each other, the data to date indicates that microbial populations are relatively unique to each type of isolated system. Nonetheless, certain trends in composition driven by physico-chemical factors and potentially biogeography, are also apparent.

Salinity

Hypersaline lakes from the Vestfold Hills (Bowman *et al.*, 2000a) and McMurdo Dry Valleys (Glatz *et al.*, 2006; Mosier *et al.*, 2007) were all dominated by *Gammaproteobacteria* and members of the *Bacteroidetes* as well as harboring lower abundance populations of *Alphaproteobacteria*, *Actinobacteria*, and *Firmicutes*. The surface waters of lakes close to marine salinity resemble marine communities dominated by *Bacteroidetes*, *Alphaproteobacteria* and *Gammaproteobacteria*, but divisions such as *Actinobacteria* and specific clades of *Cyanobacteria* have been found to be overrepresented compared to the ocean (Lauro *et al.*, 2011). Sediments from saline lakes in the Vestfold Hills (Bowman *et al.*, 2000b) and Nuramake-Ike in the Syowa Oasis (Kurosawa *et al.*, 2010)

were very similar, containing in addition to the surface clades, *Deltaproteobacteria*, *Planctomycetes*, *Spirochaetes*, *Chloroflexi* (green non-sulphur bacteria), *Verrucomicrobia* and representatives of related candidate divisions. Plankton from freshwater lakes were characterized by an abundance of *Betaproteobacteria*, although *Actinobacteria*, *Bacteroidetes*, *Alphaproteobacteria* and *Cyanobacteria* were also prominent (Pearce, 2003, 2005; Pearce *et al.*, 2005; Schiaffino *et al.*, 2009).

Trophic status

Differences in bacterial community structure are also influenced by nutrient availability. In studies of freshwater lakes in the Antarctic Peninsula and the South Shetland Islands, cluster analysis of DGGE profiles grouped together lakes of similar trophic status (Schiaffino *et al.*, 2009; Villaescusa *et al.*, 2010). Most of the variance in community structure could be explained by related chemical parameters such as phosphate and dissolved inorganic nitrogen. Similarly, three freshwater lakes, Moss, Sombre and Heywood on Signy Island are alike except that Heywood Lake is enriched by organic inputs from seals. Bacterial composition in each lake changed from winter to summer and this was again correlated to variation in physico-chemical properties (Pearce, 2005). The bacterial population of Heywood Lake had shifted from a dominance of *Cyanobacteria* towards a greater abundance of *Actinobacteria* and marine *Alphaproteobacteria* (Pearce *et al.*, 2005). This hints at a link between a copiotrophic lifestyle in the Heywood Lake *Actinobacteria* and inhibition of Antarctic freshwater *Cyanobacteria* by eutrophication. This type of study exemplifies how inferences can be made about taxa and function by examining population changes over time and over gradients of environmental parameters.

Biogeography

The relative isolation and diverse chemistries of the lakes facilitates biogeographical and biogeochemical studies. The anoxic and sulphidic bottom waters of some meromictic lakes form due to a density gradient that precludes mixing. Although sedimentation from the upper aerobic waters may occur, there is little opportunity for interchange of species with the bottom water of lakes allowing for greater divergence in community composition as nutrients can become depleted and products of metabolism can accumulate. As a result, distinct distributions of bacterial groups can inhabit these strata, and different types of microorganisms can be found in equivalent strata in different lakes.

A good example of this is the presence of common types of purple sulphur bacteria (*Chromatiales*) and green sulphur bacteria (GSB) (*Chlorobi*) in some meromictic lakes and stratified fjords in the Vestfold Hills (Burke and Burton, 1988), compared to diverse purple non-sulphur bacteria in Lake Fryxell, McMurdo Dry Valleys (Karr *et al.*, 2003). In Lake Bonney, the east and west lobes harbor overlapping but distinct communities in the suboxic waters (Glatz *et al.*, 2006). The east lobe was dominated by *Gammaproteobacteria* and the west lobe by *Bacteroidetes*, illustrating how divergent communities

can form from the same seed population. In contrast, ice communities are more readily dispersed by wind, aerosols and melt-water. 16S rRNA gene probes designed from bacteria trapped in the permanent ice-cover of Lake Bonney hybridized to microbial mat libraries sourced up to 15 km away (Gordon *et al.*, 2000). This demonstrates how a single lake may encompass microorganisms that are geographically dispersed, while also harboring others that have restricted niches and are under stronger selection pressure.

Bacterial diversity of Lake Vostok

Subglacial systems have been isolated from the open environment for hundreds of thousands to millions of years (Siegert *et al.*, 2001). The biggest of these, Lake Vostok is approximately 4 km below the continental ice-sheet. As a result they provide a reservoir of microorganisms that may have undergone significant evolutionary divergence from the same seed populations that were not isolated by the Antarctic ice cover. To date, molecular microbial studies have concentrated on the accretion ice above the ice-water interface (Priscu, 1999; Christner *et al.*, 2001). Accretion ice has been found to contain a low density of bacterial cells from *Alphaproteobacteria*, *Betaproteobacteria*, *Actinobacteria* and *Bacteroidetes* divisions closely allied to other cold environments. Molecular signatures of a thermophilic *Hydrogenophilus* species were also identified in accretion ice raising the possibility that chemoaerotrophic thermophiles were delivered to the accretion ice from hydrothermal areas in the lakes bedrock (Bulat *et al.*, 2004; Lavire *et al.*, 2006).

However, interpretation of results from samples sourced from the Lake Vostok bore hole are very challenging as it is difficult to differentiate contaminants from native Vostok microorganisms. From a study that assessed possible contaminants present in hydrocarbon-based drilling fluid retrieved from the Vostok ice core bore hole, six phylotypes were designated as new contaminants (Alekhnina *et al.*, 2007). Two of these were *Sphingomonas* phylotypes essentially identical to those found in the accretion ice-core (Christner *et al.*, 2001), which raises question about whether bacteria identified from the ice-cores are representative of Lake Vostok water, and is an example of how contamination may occur.

1.3.2 *Archaea*: methanogens and haloarchaea

Archaea have been detected mainly in anoxic sediments and bottom waters from lakes that range in salinity from fresh to hypersaline. Those with known isolates are affiliated with methanogens or haloarchaea (Bowman *et al.*, 2000b,a; Purdy *et al.*, 2003; Kurosawa *et al.*, 2010; Lauro *et al.*, 2011). Anoxia allows for the growth of methanogenic *Archaea* that mineralize fermentation products such as acetate, H₂ and CO₂ into methane, thereby performing an important step in carbon cycling. The acetoclastic methanogens thrive in environments where alternative terminal electron acceptors such as sulphate and nitrate have been depleted. One example of this is Lake Heywood where methanogenic *Archaea* were found to comprise 34% of the total microbial population in the freshwater sediment, the majority of which were *Methanosarcinales*,

which include acetate and C1-compound utilizing methanogens (Purdy *et al.*, 2003).

In general, archaeal populations appear to be adapted to their specific lake environment. Sediments from saline lakes of the Vestfold Hills were inhabited by members of the *Euryarchaeota* typically found in sediment and marine environments with the phylotypes differing between the lakes examined (Bowman *et al.*, 2000b). While a phylotype similar to *Methanosarcina* was identified, the majority were highly divergent. Similarly, *Methanosarcina* and *Methanoculleus* were detected in Lake Fryxell but other members of the *Euryarchaeota* and *Crenarchaeota* (a single sequence) were divergent, clustering only with marine clones (Karr *et al.*, 2006). Based on the lake chemical gradients and the location of these novel phylotypes in the water column the authors speculated these *Archaea* may have alternative metabolisms such as anoxic methanotrophy or sulphur-utilization.

In sediments from Lake Nurume-Ike in the Langhovde region, 205 archaeal clones grouped into three phylotypes, with the predominant archaeal clone being related to a clone from Burton Lake in the Vestfold Hills, while the other two did not match to any cultivated species (Kurosawa *et al.*, 2010). In hypersaline lakes where bottom waters do not become completely anoxic, methanogens are not present and *Archaea* have extremely low abundance. For example, only two archaeal clones of the same phylotype were recovered from deep water samples from Lake Bonney (Glatz *et al.*, 2006), and Organic Lake in the Vestfold Hills had an extremely low abundance of archaeal clones related to *Halobacteriales* (Bowman *et al.*, 2000a). In contrast to these stratified hypersaline lakes, the microbial community in the extremely hypersaline Deep Lake is dominated by haloarchaea (Bowman *et al.*, 2000a). Many of the clones identified from Deep Lake are similar to *Halorubrum* (formerly *Halobacterium*) *lacusprofundi* which was isolated from the lake (Franzmann *et al.*, 1988).

1.3.3 *Eucarya* perform multiple ecosystem roles

Single-celled *Eucarya* are important members of Antarctic aquatic microbial communities. In many Antarctic systems, eucaryal algae are the main photosynthetic organisms and in others, only heterotrophic protists occupy the top trophic level. As eucaryal cells are generally large with characteristic morphologies, microscopic identifications have been used. However, microscopy is unable to classify smaller cells such as nanoflagellates with high resolution, although these may constitute a high proportion of algal biomass. For example, five morphotypes of *Chrysophyceae*, evident in Antarctic lakes were unidentifiable by light microscopy but were able to be classified using DGGE and DNA sequencing (Unrein *et al.*, 2005). Consistent with this, molecular studies specifically targeting eucaryal diversity (Unrein *et al.*, 2005; Mosier *et al.*, 2007; Bielewicz *et al.*, 2011) have identified a much higher level of diversity than previously suspected, and the studies have discovered lineages not previously known to be present such as silicoflagellates of the family *Dictyochophyceae* (Unrein *et al.*, 2005) and fungi (Mosier *et al.*, 2007; Bielewicz *et al.*, 2011).

Most *Eucarya* in Antarctic lakes are photosynthetic microalgae that are present

in marine environments with a wide distribution including chlorophytes, haptophytes, cryptophytes and bacillariophytes. Molecular methods have afforded deeper insight into the phylogenetic diversity within these broader divisions and have revealed some patterns in their distribution. Using 18S rRNA gene amplification and DGGE, the same chrysophyte phylotypes were identified in lakes from the Antarctic Peninsula and King George Island despite being 220 km apart (Unrein *et al.*, 2005) indicating these species may be well-adapted to Antarctica or highly dispersed. Similarly, an unknown stramenopile sequence was detected throughout the 18S rRNA clone libraries of Lake Bonney demonstrating a previously unrecognized taxon occupied the entire photic zone in the lake (Bielewicz *et al.*, 2011). In contrast, other groups showed distinct vertical and temporal distributions with cryptophytes dominating the surface, haptophytes the midwaters and chlorophytes the deeper layers during the summer while stramenopiles increased in the winter (Bielewicz *et al.*, 2011).

The influence of flagellates on ecosystem function is not necessarily clear-cut as they can simultaneously inhabit several trophic levels. For instance, in Ace Lake the mixotrophic phytoflagellate *Pyramimonas gelidocola* derives a proportion of its carbon intake through bacterivory (Bell and Laybourn-Parry, 2003) but in the nearby Highway Lake, it uptakes dissolved organic carbon (Laybourn-Parry *et al.*, 2005). This illustrates potential limitations for deriving ecosystem level functions from taxonomic studies alone, even with taxa that appear physiologically straightforward. Further studies are necessary to determine the basis for apparent specific adaptations of some species to particular lakes or lake strata, and for the cosmopolitan distribution of others. Here, molecular based research of the kind that has been applied to bacteria such as functional gene surveys will undoubtedly help answer these questions.

1.4 Integrative studies to derive whole ecosystem function

The relatively low diversity of Antarctic microbial food-webs existing within effectively closed systems allows for an integrative understanding of the microbial community and biogeochemical cycling to be obtained. This can be achieved by combining molecular information of the taxonomic or functional genes with abiotic parameters or reaction rates.

1.4.1 A single gene approach

One such study was conducted on Blood Falls, an outflow of anoxic ferrous brine from the Taylor Glacier in the McMurdo Dry Valleys, where an unusual iron–sulphur cycle was inferred to exist. The water is sulphate-rich, exists in permanent darkness and is estimated to have been isolated from external inputs for 1.5 million years (Mikucki *et al.*, 2009). PCR-screening for SSU and functional gene markers for dissimilatory sulphur conversions was conducted to piece together the ecosystem function. 16S rRNA gene analysis showed the community was dominated by a close relative of *Thiomicrospira arctica*, an autotrophic sulphur-oxidizing member of the *Gammaproteobacte-*

ria (*Thiotrichales*), as well as sequences related to *Delta-* and *Gammaproteobacteria* capable of iron and/or sulphur compound reduction, and *Bacteroidetes* capable of heterotrophic growth on organic compounds (Mikucki and Priscu, 2007). A large proportion of adenosine 5'-phosphosulphate reductase genes related to those involved in dissimilatory sulphate metabolism were detected. However, dissimilatory sulphite reductase (*dsrA*) was not present and radioisotope data indicated sulphide is not produced. The implication of this is that sulphate reduction does not proceed to sulphide as typically occurs in other aquatic systems, and instead sulphate is expected to be regenerated via an alternative cycle with Fe(III) acting as the terminal electron acceptor (Mikucki *et al.*, 2009). This is a fascinating example of how a closed system has adapted to sustain life in the absence of light energy through the use of atypical chemical cycling; a pathway that was speculated to have possibly occurred in the ancient Neoproterozoic ocean (1,000 to 500 mya) (Mikucki *et al.*, 2009).

1.4.2 ‘-omics’ approaches

Motivation for adopting metagenomic and other ‘-omics’ approaches has stemmed from the inability to obtain physiological data from PCR-based taxonomic surveys and limited scope offered by screening of single functional gene markers. Although metabolic capacity can often be inferred from diversity of a single gene if there are cultured representatives with the detected genes, close relatives with defined physiologies are often not available. Inferring function can be problematic due to the existence of species and even strain level differences. For example, the majority of *Gammaproteobacteria* that have been detected in hypersaline lakes are relatives of *Marinobacter* suggesting that this genus is particularly adapted to hypersaline systems (Bowman *et al.*, 2000a; Glatz *et al.*, 2006; Matsuzaki *et al.*, 2006; Mosier *et al.*, 2007). Nonetheless, *Marinobacter* species from different lakes appear biochemically distinct as isolates from hypersaline lake Suribati-Ike were all able to respire dimethylsulphoxide (DMSO) but not nitrate (Matsuzaki *et al.*, 2006). In contrast, those from the west lobe of Lake Bonney were all able to respire nitrate (Ward and Priscu, 1997). Interestingly, in the east lobe of the same lake, nitrate respiration was inhibited although a near-identical *Marinobacter* phylotype was present; it was speculated that the inhibition may have been caused by an as yet unidentified chemical factor (Ward *et al.*, 2005; Glatz *et al.*, 2006). Metagenomics presents itself as a powerful tool to explore microbial communities because it can provide extensive information on the genetic content of uncultured taxa that is embedded within a whole community genomic context.

The first large scale metagenomic study of an Antarctic lake specifically targeted viruses (López-Bueno *et al.*, 2009). As obligate parasites, viruses cannot be isolated without first having a susceptible host in culture, which curtails the possibility of surveying anything but defined viral populations through culture-based techniques. Viruses do not have a universal gene that may be used as a taxonomic marker, or even universal genetic material, so metagenomic sequencing is an ideal approach for surveying viral diversity. Analysis of the viral component of the freshwater Lake Limnopolar,

Livingston Island uncovered the greatest depth of viral diversity of any aquatic system to date (López-Bueno *et al.*, 2009). Representatives from 12 viral families were detected, but unlike the two previous viromes that had been published at that time using comparable techniques, ssDNA viruses and large dsDNA viruses that putatively infect *Eucarya* were the dominant viral types. The ssDNA viruses were related to circoviruses, geminiviruses, nanoviruses and satellites; viruses previously only known to infect plants and animals indicating they are much more diverse than previously suspected and may constitute new viral families. Samples taken in summer showed a shift in the viral community composition towards phycodnaviruses similar to *Ostreococcus tauri* virus, OtV5. This shift potentially reflects an increase in the host algae that are stimulated to bloom by the increased light availability. Subsequent analysis has indicated multiple displacement amplification of low quantities of starting DNA used in virome study leads to a stochastic amplification bias rendering the metagenomes non-quantitative (Yilmaz *et al.*, 2010). Nonetheless, clear is that viruses perform a crucial role in shaping community structure, driving host evolution and contributing to the dissolved nutrient pool (Danovaro *et al.*, 2011) and this pioneering study has shed light on their diversity in the Antarctic environment.

A combined metagenomic and metaproteomic approach has been applied to the study of Ace Lake in the Vestfold Hills (Ng *et al.*, 2010; Ng, 2010; Lauro *et al.*, 2011). From this study, the main carbon, nitrogen and sulphur cycles within the stratified water column of Ace Lake were able to be described (Lauro *et al.*, 2011). A strictly interdependent sulphur cycle was found to exist between GSB at the oxycline and sulphate-reducing bacteria (SRB) in the anaerobic zone. The GSB were able to convert sulphide to sulphate but lacked assimilatory sulfate reduction capability (Ng *et al.*, 2010; Ng, 2010; Lauro *et al.*, 2011). Enzymes involved in dissimilatory sulphate reduction (adenylylsulphate reductase and dissimilatory sulphite reductase) were detected in the metaproteome of the anaerobic zone thereby completing the dissimilatory sulphur cycle (Ng *et al.*, 2010; Ng, 2010). It was also apparent that short circuiting of the nitrogen cycle was occurring. Nitrogen fixation proteins were not found in the metaproteome, likely due to preferential assimilation of ammonia (Ng *et al.*, 2010; Ng, 2010; Lauro *et al.*, 2011). Genes involved in nitrification and denitrification were underrepresented indicating the system had shifted away from nitrogen mineralization consistent with a mechanism to conserve bioavailable nitrogen (Lauro *et al.*, 2011). It is noteworthy that these findings are in contrast to lakes in the Dry Valleys where ammonia monooxygenase genes were present in the aerobic zone of six lakes of various salinities and extents of stratification (Voytek *et al.*, 1999). Overall, strong depth partitioning, loss of biochemical pathways common in other aquatic systems and the selection of key species were identified as important steps in transiting the microbial population from marine to a meromictic lake ecosystem.

Further results from the study of Ace Lake are presented in Chapter 2 and metagenomic and metaproteomic analyses of Organic Lake is described in Chapters 3 and 4.

1.5 Objectives

The overall aim of this thesis was to explore the microbial communities of Antarctic lakes from an ecosystem level perspective by combining metagenomics, metaproteomics and physico-chemical data. Ace Lake and Organic Lake, two lakes in the Vestfold Hills, were chosen as the study sites as there are extensive historic environmental records available for these lakes. As meromictic lakes, differences in the microbial population were able to be examined along the vertical gradients within the lakes. Finally, as they are marine-derived systems, comparison between the lake system and the marine environment was used to identify specific lake adaptations.

The specific objectives of the research were:

1. Develop epifluorescence microscopy and metaproteomic methods to complement to metagenomic sequencing.
2. Determine the microbial and viral composition of Antarctic lake communities, their functional potential and infer the ecological roles of populations in the community.
3. Integrate environmental and biological data to model the lake microbial interactions and biogeochemical processes.

Chapter 2

Development of methods to complement metagenomic sequencing for an integrative study of Ace Lake

Co-authorship statement

Sections from this chapter have been published as:

Federico M. Lauro, Matthew Z. DeMaere, **Sheree Yau**, Mark V. Brown, Charmaine Ng, David Wilkins, Mark J. Raftery, John A.E. Gibson, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffrey M. Hoffman, Torsten Thomas, and Ricardo Cavicchioli. An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME Journal* 5:879–895, 2011.

I performed the epifluorescence microscopy, metaproteomic mass spectral analysis and drafted the corresponding sections of the publication.

Contributions by others that support the work presented in this chapter are as follows. Research was designed by Federico Lauro, Mark Brown, Torsten Thomas, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown, Torsten Thomas, Jeffrey Hoffman and Ricardo Cavicchioli. DNA extraction and clone library preparation was performed by Cynthia Andrews-Pfannkoch and Jeffery Hoffman. DNA sequencing quality control was performed by Matthew Lewis. Metagenomic sequence filtering, mosaic assembly and annotation was performed by Matthew DeMaere. Protein extraction, one-dimensional sodium dodecyl sulphate-polyacrylamide gel electrophoresis and liquid chromatography mass spectrometry and preliminary analysis performed by Charmaine Ng. Assistance in mass spectrometry was provided by Mark Raftery.

2.1 Abstract

Ace Lake is a saline meromictic lake and the most studied lake in the Vestfold Hills, Antarctica. As a system of moderate biological complexity with extensive historic physical and chemical data, it was chosen as a site to implement an integrative study of the lake ecosystem. Metagenomic analysis of Ace Lake revealed microbial taxa and metabolic genes were stratified according to the lake's water column structure and also was able to infer potential for nutrient cycling (Lauro *et al.*, 2011). This study aimed to generate independent datasets complementary to metagenome data as part of the integrative analysis of the lake. A method for visualising and enumerating cells and virus-like particles (VLPs) using epifluorescence microscopy was developed that does not require use of the relatively expensive Anodisc filters. Microscopic examination confirmed the efficacy of the sequential filtration procedure used to size fractionate the lake microbiota and determined the densities of cells and VLPs. Furthermore, it independently verified the lack of VLPs associated with the lake's green sulphur bacteria (GSB). Previous metaproteomic analysis of the Ace Lake depth profile using cross-species matching of mass spectra yielded few protein identifications (Ng, 2010). However, metaproteomic analysis of the GSB layer using matching GSB metagenomic sequences as the search database gave a 3-fold increase in protein identifications (Ng *et al.*, 2010; Ng, 2010). A metaproteomic analysis workflow was developed that achieved substantial improvements in protein identification rates by similarly using metagenomic databases matched to the metaproteomic samples. This involved application of the software package SCAFFOLD 2.0 to mass spectral analysis, which additionally allowed for protein abundance estimates by spectral counting. Functional protein groups were identified that were overrepresented in each zone of the lake. These data proved crucial to support a comprehensive description of the entire Ace Lake ecosystem.

2.2 Introduction

Ace Lake is a meromictic saline lake in the Vestfold Hills that separated from the sea ~5,000 BP (Bird *et al.*, 1991). Extensive physical, chemical and biological data has been collected from Ace Lake since 1974 (see <https://data.aad.gov.au/aadc/lakes/>). The system is microbially-dominated and has reduced species diversity (Bowman *et al.*, 2000b) with the only metazoan life present being callanoid copepods. Ace Lake is a highly stratified lake system that is 25 m at its deepest point (Figure 2.1). It is ice-covered for ~11 months of the year and generally thaws in January (Rankin *et al.*, 1999). Water is marine-derived and a largely neutral water balance has ensured salinity is close to that of seawater. The lake is physically separated into an aerobic mixolimnion, a steep chemo/oxycline at 12.7 m and an anoxic monimolimnion. The monimolimnion is sulfidic and methanogenic; both compounds have presumably accumulated through the activity of sulphate-reducing bacteria (SRB) and methanogenic archaea respectively (Rankin *et al.*, 1999; Lauro *et al.*, 2011) (Figure 2.1). As a physically and chemically well-characterised system of moderate diversity, Ace Lake was chosen as a model ecosys-

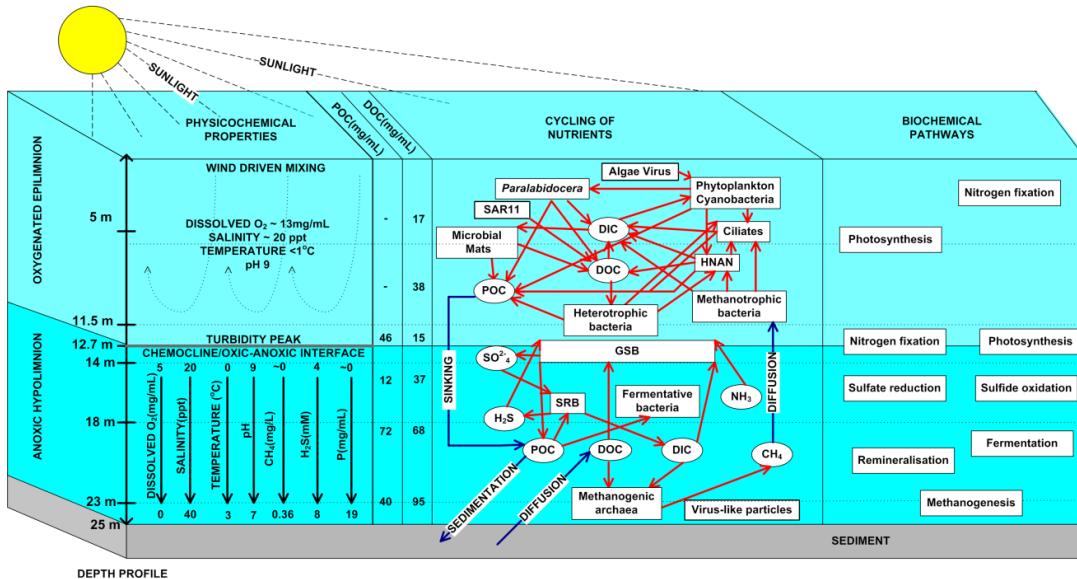


Figure 2.1: Physico-chemical and biological structure of Ace Lake. In the panel describing nutrient cycling biotic (red arrows) and abiotic (blue arrows) are shown. The panel describing biochemical pathways shows the labels of the pathways at the depths where they are most significant. The figure is from Lauro *et al.* (2011) with some modifications.

tem to implement a whole systems level analysis to piece together ecosystem functioning. Samples were obtained from the mixolimnion (5 and 11.5 m), the chemo/oxycline (12.7 m) and the monimolimnion (14, 18 and 23 m). Sampling was conducted according the design of the global ocean sampling (GOS) expedition (Rusch *et al.*, 2007) by using size fractionation of microbial biomass onto 3.0, 0.8 and 0.1 μm membrane filters.

From the metagenomic analysis conducted on the Ace Lake samples, significant differences were found in taxonomic composition between each size fraction and between the three zones of Ace Lake (Lauro *et al.*, 2011). The mixolimnion community is similar to a marine surface water assemblage consisting of a high abundance of the SAR11 clade of *Alphaproteobacteria* related to “*Candidatus Pelagibacter ubique*” and green algae of the order *Mamiellales*. However, diversity is reduced by one order of magnitude (Lauro *et al.*, 2011). Unlike Southern Ocean surface water, the mixolimnion is over-represented in *Cyanobacteria* related to *Synechococcus* and *Actinobacteria*, which may represent taxa that mark the transition from a marine to lake community. A dense, near-clonal population of green sulphur bacteria (GSB) related to *Chlorobium* termed C-ace reside at the chemo/oxycline at 12.7 m (Ng *et al.*, 2010; Lauro *et al.*, 2011). Below, in the monimolimnion is a highly diverse community that includes SRB and methanogenic Archaea. The viral community comprises *Phycodnaviridae*, *Myoviridae*, *Siphoviridae*, *Podoviridae* and unidentified viral families (Lauro *et al.*, 2011). Bacteriophage sequences were abundant in the bottom waters, whereas the surface community, was dominated by phycodnaviruses (Lauro *et al.*, 2011).

Metagenomics is an extremely powerful tool for interrogating microbial communities. However, a systems-level understanding ultimately requires the integration of data

which metagenomic sequencing alone does not provide such as abiotic parameters and functional activity (Handelsman, 2008; Warnecke and Hugenholtz, 2007). The focus of this study was to provide datasets to complement metagenome data.

A modified epifluorescence microscopy procedure was developed to determine cellular and viral densities and validate the efficacy of size-fractionation. Development of a revised method was necessary due to inability to source 25 mm diameter 0.02 µm pore-size Anodisc filters (Whatman) that have long been used with fluorescent nucleic acid dyes for this purpose (Hennes and Suttle, 1995; Noble and Fuhrman, 1998; Patel *et al.*, 2007). They were marked for discontinuation in December 2008 after the take-over of Whatman by GE Healthcare, and was the cause of a global shortage that was strongly opposed by the viral ecology community (Torrice, 2009). Since conducting this research, supply of 25 mm Anodisc filters has resumed, albeit at much greater cost per filter. The need for alternative methodologies has been great enough that protocols were developed independently by other research groups (Budinoff *et al.*, 2011; Diemer *et al.*, 2012) stressing the utility of alternatives.

Identification of proteins from a microbial community (metaproteomics) indicates which populations or processes are active in the environment and is thus a powerful tool for understanding ecosystems. Preliminary metaproteomic analysis along the vertical profile of Ace Lake has been performed using a cross-species genomic database comprising the National Center for Biotechnology Information (NCBI) non-redundant database (NR) (Ng, 2010). A focussed metaproteogenomic analysis conducted on the dense GSB layer using the matched GSB metagenome as the database resulted in 3-fold increase in protein identifications compared to using the NR database (Ng, 2010). This indicates large gains in metaproteome coverage are possible using a matched metagenomic database (Ng, 2010). However, protein identification by the ‘shotgun’ proteomics approach favoured in metaproteomics (Ram *et al.*, 2005) is computationally challenging when dealing with samples of higher species diversity. In this study, a bioinformatic analysis workflow was devised tailored to Ace Lake metaproteomic data to identify proteins as well as estimate their abundance.

2.3 Materials and methods

2.3.1 Ace Lake samples

Water samples were collected from Ace Lake ($68^{\circ}28'23.2''S$, $78^{\circ}11'20.8''E$), Vestfold Hills, Antarctica on 21 and 22 December 2006 as described previously (Ng *et al.*, 2010; Ng, 2010; Lauro *et al.*, 2011). Briefly, a hole positioned above the deepest point (25 m depth) of the lake was drilled through the 2 m ice cover of Ace Lake to reach the lake surface. Microbial biomass was collected by sequential size fractionation through a 20 µm pre-filter directly onto 3.0, 0.8 and 0.1 µm pore-size 293 mm polyethersulfone membrane filters (Rusch *et al.*, 2007), along the depth profile. Two independent sets of filters were obtained, one set for metagenomics and one for metaproteomics.

A sonde probe (YSI model 6600, YSI Inc., Yellow Springs, OH, USA) was used

Table 2.1: Summary of metagenomic data for Ace Lake December 2006 profile. S, Sanger sequencing data; *Scaffolds from 0.1 μ m fraction were assembled from a hybrid of Sanger and 454 sequences.

| ID | Depth (m) | Size (μ m) reads | Trimmed reads | ORFs (reads) | Annotated COG/KEGG ORFs | >10 kbp scaffolds (reads) | Annotated scaffold ORFs |
|------------|-----------|----------------------------|--------------------|-----------------------------------|---------------------------------|---------------------------|-------------------------|
| GS232 5 | 0.1 | S: 281,490 454: 539,536 | 421,252 638,757 | 112,490/96,771 109,551/103,201 | * | * | |
| | 0.8 | 454: 468,122 | 485,021 | 150,660/135,451 | 2,809 (349,015) 269 (66,743) | 45,281 3,215 | |
| | 3.0 | 454: 160,835 | 138,191 | 24,920/22,240 | 33 (2,980) | 353 | |
| | | | | | | | |
| GS231 11.5 | 0.1 | S: 283,663 | 427,889 | 124,332/107,523 | * | * | |
| | 0.8 | 454: 523,650 | 608,671 | 99,175/95,266 | 2,814 (390,490) | 47,987 | |
| | 3.0 | 454: 474,419 | 511,909 | 218,126/176,332 | 174 (161,891) | 2,321 | |
| GS230 12.7 | 0.1 | S: 54,446 | 75,576 | 42,790/41,391 | * | * | |
| | 0.8 | 454: 442,389 | 492,995 | 201,203/227,726 | 88 (282,232) | 3,039 | |
| | 3.0 | 454: 529,711 | 555,328 | 209,078/234,682 | 86 (313,550) | 2,187 | |
| | | | | | | | |
| GS229 14 | 0.1 | S: 10,042 | 14,326 | 5,261/4,469 | * | * | |
| | 0.8 | 454: 413,992 | 458,942 | 100,045/88,300 | 228 (22,556) | 2,443 | |
| | 3.0 | 454: 453,205 | 435,534 | 142,743/129,403 | 139 (45,083) | 2,118 | |
| GS228 18 | 0.1 | S: 9,672 | 15,077 | 3,667/3,008 | * | * | |
| | 0.8 | 454: 362,490 | 389,077 | 51,312/44,290 | 29 (1,815) | 260 | |
| | 3.0 | 454: 544,302 | 556,243 | 186,455/163,878 | 154 (14,806) | 1,334 | |
| GS227 23 | 0.1 | S: 100,085 | 160,302 | 33,462/27,302 | * | * | |
| | 0.8 | 454: 482,527 | 547,170 | 84,257/73,074 | 1,136 (51,163) | 12,339 | |
| | 3.0 | 454: 553,234 | 611,717 | 161,973/137,632 | 105 (7,904) | 825 | |
| | | | | | | | |
| TOTAL | - | 8,103,379 | 8,926,759 | 2,487,574/2,283,749 | 8,269 (1,771,149) | 126,585 | |

to record depth, dissolved oxygen content, pH, salinity, temperature and turbidity throughout the water column of the lake. Total organic carbon was determined using a total organic carbon analyzer, TOC-5000A (Shimadzu, Kyoto, Japan) equipped with a ASI-5000A auto sampler (Shimadzu), and particulate organic carbon by standard protocols (<http://www.epa.gov/glnpo/lmmmb/methods/about.html>) at the Centre for Water and Waste Technology, UNSW.

2.3.2 DNA extraction, sequencing and data cleanup

DNA extraction and Sanger sequencing was performed on 3730xl capillary sequencers (Applied Biosystems, Carlsbad, CA, USA) and pyrosequencing on GS20 FLX Titanium (Roche, Branford, CT, USA) at the J. Craig Venter Institute in Rockville, MD, USA (Rusch *et al.*, 2007). The scaffolds and annotations will be available via Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) and public sequence repositories such as the NCBI and the reads will be available via the NCBI Trace Archive.

Sanger reads and pyrosequencing reads were trimmed and filtered as described in Lauro *et al.* (2011). See Table 2.1 for a summary of the metagenomic data and sample IDs.

2.3.3 Metagenomic DNA assembly and annotation

Mosaic assemblies were generated for each sample fraction using Celera whole genome shotgun (WGS) Assembler v5.3 (Myers *et al.*, 2000) as described previously (Ng *et al.*, 2010; Lauro *et al.*, 2011). Each 0.1 µm fraction assembly was a hybrid of Sanger and 454 read data. See Table 2.1 for a summary of metagenomic assemblies. Annotation of each sample fraction assembly was carried out using an in-house pipeline described in DeMaere *et al.* (2011), wherein the pipeline stages consisted of genomic feature detection and subsequent annotation by basic local alignment search tool (BLAST) comparison to NR, Swissprot and Kyoto Encyclopedia of Genes and Genomes (KEGG)-peptide sequence databases and by biosequence analysis using profile hidden Markov models (HMMER) comparison against the Institute of Genomic Research curated protein database (TIGRFAM) (Haft *et al.*, 2001), clusters of orthologous groups (COG) (Tatusov *et al.*, 1997, 2003) and known marker genes (von Mering *et al.*, 2007).

2.3.4 Epifluorescence microscopy

Samples of unfiltered lake water and the flow-through from 3.0 and 0.8 µm filters from all depths were collected on November 2008 and fixed on site in formalin 1% (v/v). The samples were stored at –80°C for subsequent direct counts of cells and virus-like particles (VLPs). Enumeration was performed according to the method of Patel *et al.* (2007) with modifications. Lake water samples were filtered onto 0.01 µm pore-size polycarbonate filters (25 mm Poretics, GE Osmonics, Minnetonka, MN, USA). Filters were air dried, then placed with the back of the filter on top of a 30 ml aliquot of 0.1% (w/v) molten low-gelling-point agarose and allowed to dry at 30°C. Samples were stained by the addition of 2 µl working solution (1 in 400 dilution in 0.02 µm filtered sterile Milli-Q) of SYBR Gold (Molecular Probes, Eugene, OR, USA) to 20 µl of mounting medium (VECTASHIELD HardSet, Vector Laboratories, Burlingame, CA, USA). Stained samples were counted immediately. Samples were visualised under wide-blue filter set (excitation 460–495 nm, emission 510–550 nm) with an epifluorescence microscope (Olympus BX61, Hamburg, Germany).

2.3.5 Protein extraction

Proteins were extracted 0.1 µm pore-size membrane filters from the six depths (5, 11.5, 12.7, 14, 18 and 23 m) as described previously (Ng *et al.*, 2010; Ng, 2010; Lauro *et al.*, 2011). Briefly, separate extractions were performed on quarters of the 0.1 µm filters which served as technical replicates. Cells were lysed by three freeze-thaw cycles in liquid nitrogen and sonication in lysis buffer containing 10 mM Tris-EDTA (pH 8), 20 µl of protease inhibitor cocktail VI (Calbiochem), 0.1% sodium dodecyl sulphate, and 1 mM dithiothreitol. Buffer exchange into 10 mM Tris-EDTA (pH 8) and concentration of soluble extracted proteins was performed in a 5 kDa cut-off Amicon Ultra-15 filter unit (Millipore).

2.3.6 1D-SDS PAGE and LC-MS-MS

Extracted proteins were separated by one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis (1D-SDS PAGE) containing 12% SDS using a Mini-PROTEAN system (Bio-Rad, Sydney, NSW, Australia) as described previously (Saunders *et al.*, 2006). Trypsin digestion, liquid chromatography (LC) and mass spectrometry was conducted as described previously (Ng *et al.*, 2010; Ng, 2010; Lauro *et al.*, 2011). Briefly, gel slices were subject to a series of reduction, alkylation and dehydration reactions before digestion of proteins with trypsin. Peptides were separated by high performance liquid chromatography (HPLC) and positive ions generated by electrospray for mass spectra acquisition by the LTQ-FT Ultra mass spectrometer. Peak lists were generated using EXTRACT_MSM from MASCOT DAEMON (Matrix Science, Thermo, London, UK) using the default parameters.

2.3.7 Metaproteomic mass spectra analysis

The spectra generated from mass spectrometry (MS) were searched against the protein sequence database corresponding to that depth constructed from the 0.1 µm mosaic assemblies using MASCOT version 2.1 (Matrix Science). MASCOT DISTILLER (Applied Biosystems) was used as the data import filter with the following criteria applied to the two dimensional mass spectrometry (MS-MS) ion search: a maximum of one missed cleavage for trypsin, peptide mass tolerance of \pm 4 ppm, a fragment mass tolerance of \pm 0.6 Da and variable modifications of acrylamide, carbamidomethyl and oxidation. The number of protein sequences in each database were as follows: 5 m, 138,208; 11.5 m, 133,948; 12.7 m, 27,142; 14 m, 62,436; 18 m, 71,512; and 23 m, 128,878. SCAFFOLD 2.0 (version Scaffold_2.05.01, Proteome Software Inc., Portland, OR, USA) was used to validate MS-MS-based peptide and protein identifications. Peptide and protein identifications were accepted if they could be established at >95% and 99% probability, respectively, as specified by the PEPTIDE PROPHET algorithm (Keller *et al.*, 2002). Protein identifications required the identification of at least two peptides.

Proteins that contained similar peptides and could not be differentiated based on MS-MS analysis alone were grouped to satisfy the principles of parsimony and are referred to as a protein group. Spectral counting was used to semi-quantitatively estimate protein abundance. The total assigned spectra that matched to each identified protein were exported from SCAFFOLD 2.0. For similar proteins that have shared peptides (a protein ambiguity group), spectra were assigned to the protein with the most unique spectra. To normalise for variation in total spectra acquired between sample replicates, the number of spectra of each protein was multiplied by the average total spectra divided by the total spectra of the individual replicate. The spectral count of each protein was averaged across the replicates. As longer proteins are more likely to be detected, the average spectral counts were divided by the length of the protein. This value is equivalent to the normalised spectral abundance factor (Florens *et al.*, 2006; Zybailev *et al.*, 2006). In order to compare the relative abundance of proteins between depths, the normalised spectral abundance factor was divided by the average read depth

of the contig (scaffold or degenerate) to which the protein mapped.

If >90% of a scaffolds length consisted of surrogate (highly degenerate unitig) sequence, the average read depth of the surrogate was used. For identified proteins that were part of a protein group the longest protein length and largest read depth value in the group was used. Pairwise comparisons of each zone were conducted on COG assigned proteins. The normalised spectral counts from each protein was aggregated based on their COG annotation. All proteins that were part of an ambiguity group were confirmed to share the same COG annotation to ensure counts were not biased because of the common spectra.

The summed spectral counts from 5 and 11.5 m (mixolimnion), and 14, 18 and 23 m (monimolimnion) were pooled. Statistical significance of differences between each zone was assessed using Fisher's exact test, with confidence intervals at 99% significance calculated by the NewcombeWilson method and Holm-Bonferroni correction (p-value cutoff of 1e-5) in Statistical Analysis of Metagenomic Profiles (STAMP) (Parks and Beiko, 2010). All proteins identified, including their gene identifier, normalised spectral abundance, COG and KEGG Orthology identifiers, KEGG locus tag and matching COG or KEGG description are provided in Appendix B.

2.4 Results and discussion

2.4.1 Development of an epifluorescence microscopy method

An epifluorescence microscopy method was developed to allow examination of cellular morphology and enumeration of cells and VLPs in Ace Lake. The standard method for cell and VLP counts uses 25 mm diameter 0.02 μm pore-size Anodisc filters (Patel *et al.*, 2007). As mentioned above, the supply of these filters were discontinued in December 2008 (Torrice, 2009), which necessitated development of an alternative protocol.

Clear polycarbonate Track Etch (PCTE) membrane filters were selected for use as a viable alternative product as they have a defined pore-size of sufficiently small diameter (0.01 μm) to capture VLPs. Furthermore, they have been used in earlier studies for VLPs enumeration (Hara *et al.*, 1991; Proctor and Fuhrman, 1992), have a long history of use with the enumeration of cells (Hobbie *et al.*, 1977) and therefore require no new materials to be easily adopted for use. Finally, use of PCTE membranes is approximately 10 times cheaper than the use of Anodiscs and are not subject to drops in availability. However, PCTE membranes have several reported shortcoming that have precluded their standard use for viral enumeration. The 0.01 μm 25 mm PCTE filters are difficult to handle compared to 25 mm Anodisc filters that have a plastic support ring around the edge. Also, PCTE filters appear to have higher background fluorescence, have a slow flow-rate taking ~1–1.5 hours to filter 2 ml (Hara *et al.*, 1991) and have been reported to give VLP counts an order of magnitude lower than that of Anodiscs (Budinoff *et al.*, 2011).

The protocol used is detailed in section 2.3 *Materials and methods* where the main challenges of using PCTE were overcome for the purposes required for this study. 0.01

μm pore-size PCTE filters can form creases or wrinkles when mounted in the vacuum filter holder. Good placement of the PCTE filter was achieved by touching the edge of the PCTE filter against the damp backing filter, making sure it was aligned and then gently and evenly releasing it in a single direction. $0.01\ \mu\text{m}$ pore-size PCTE filters also easily become statically charged and will become attracted to surfaces so careful handling is required during manipulation. The greatest drawback in the use of PCTE filters of such small pore-size is they similarly have a tendency to crinkle when mounted on the glass slide that can make visualisation of cells and VLPs on a single focal plane difficult. Agarose was used to embed the filters to help flatten the membrane and aid in mounting. However, this was not strictly necessary if filters are dried well and pressed flat against the glass slide with a minimal volume of mountant. Even though careful handling is possible, this method likely requires more technical replicates than Anodisc filters because any localised regions with VLPs outside the focal plane will not be counted leading to greater deviation in counts. Furthermore, as similarly reported by Diemer *et al.* (2012) for $0.03\ \mu\text{m}$ PCTE membranes, distribution of VLPs on the membrane appears more variable than for Anodiscs with local regions devoid of VLPs and others with pooling of VLPs. The patchier VLPs distribution was attributed to greater irregularity in pore distribution compared to that of Anodisc filters and was considered to be the main contributing factor to variability in VLP counts (Diemer *et al.*, 2012)

Background fluorescence was at an acceptable level when the SYBR Gold stain is only incorporated into the mountant after filtration rather than staining in the column. This is one key difference between this protocol and others that use PCTE membranes (Hara *et al.*, 1991; Proctor and Fuhrman, 1992; Diemer *et al.*, 2012). Other protocols have used $0.08\ \mu\text{m}$ membranes pre-stained with Irgalan Black (Proctor and Fuhrman, 1992) or $0.03\ \mu\text{m}$ pre-stained with Sudan Black B (Diemer *et al.*, 2012) to minimise background fluorescence. Pre-staining of $0.01\ \mu\text{m}$ membranes would be an option for future optimisation of this protocol.

Only one prior report of filtration of natural water onto the PCTE membranes with $0.01\ \mu\text{m}$ pore-size was found (Hara *et al.*, 1991). Such a small pore-size necessitates a very strong seal of the filter column against the glass base with no air leaks. Leakage was eliminated by sealing the fritted base to the column with laboratory film. However, the slow flow-rate was not a property of these PCTE filters that could be overcome. As the filtration and visualisation was performed on fixed samples in the laboratory rather than in the field, the time taken for filtration of 2–3 hours for each sample was not deemed problematic for the purposes of this study. Counts of VLPs have been shown to decrease dramatically with time even when preserved with aldehyde-based fixatives (Wen *et al.*, 2004). Collecting larger sample volumes and flash freezing in liquid nitrogen and storing at -80°C helps to slow this effect (Patel *et al.*, 2007). Due to logistic constraints of working in the Antarctic, counts in the field were not possible so VLP counts are expected to be underestimated. However, all samples were processed within a similar time frame so the relative VLP abundances down the depth profile and

between size fractions are expected to be accurate as all samples would have undergone comparable amounts of VLP loss.

Overall, a viable alternative method was successfully developed for visualisation and enumeration of cells and VLPs using PCTE membrane filters for use in this study (Figure 2.2). To be a competitive alternative to Anodisc filters in terms of enumeration accuracy, further work is required that compares counts using this method and the current standard Anodisc-based protocol (Patel *et al.*, 2007) with viral samples of known densities.

2.4.2 Size and depth stratification of the community supported by cell and VLP densities

Epifluorescence microscopy images of Ace Lake microbiota showed a clear decrease in larger and particularly filamentous cells between filter size fractions (Figure 2.2). The sequential filtration of suspended microbial biomass from aquatic environments has been utilised as part of the landmark Sargasso Sea metagenomic study (Venter *et al.*, 2004) and GOS expedition (Rusch *et al.*, 2007). The Ace Lake samples were collected using the same sampling strategy as the GOS dataset but has sequence information from all three filter sizes and was able determine if size fractionation is effective at separating the microbial community and if functional differences are associated with size. Both taxonomic and functional gene composition were shown in the metagenomic analysis to differ significantly with size fraction and is indicative of resource partitioning in the community (Lauro *et al.*, 2011). The clear difference between the size fractions indicated the sequential filtration process is an effective means to separate the community and that different sized populations in the community have different ecological roles. For example, *Flavobacteria* were only detected in the 3.0 μm size fraction mixolimnion metagenomes and were inferred to be involved in mineralisation of particulate matter (Lauro *et al.*, 2011). This correlates with the absence of long rod shaped cells in the >0.8 μm microscopic image (Figure 2.2) indicating these larger cells are likely to be copiotrophic *Flavobacteria* involved in degradative processes. Depth was another variable which strongly drove differences in the Ace Lake community (Lauro *et al.*, 2011). This was again evident in the microscopy images that show the appearance of long filamentous cells only in the 12.7 m and monimolimnion samples (Figure 2.2).

Cell and VLP densities are not obtainable from metagenomic sequence data and necessitates an independent method of determination. The first studies to determine viral densities in natural systems found that viruses are the most numerous biological entities on the planet and likely play a large role in plankton mortality in the ocean (Bergh *et al.*, 1989; Proctor and Fuhrman, 1990). VLP counts from marine environments vary with depth and trophic status ranging from 10^6 to 10^8 VLP ml^{-1} (Suttle, 2005). Both cellular and VLP counts are linked with environmental factors such as trophic status (Lauro *et al.*, 2009). Enumeration of cells and VLPs in Ace Lake showed cell densities were lowest in the relatively oligotrophic mixolimnion ($0.8\text{--}1.3 \times 10^6$ cell ml^{-1}) and were by an order of magnitude higher in the copiotrophic monimolimnion

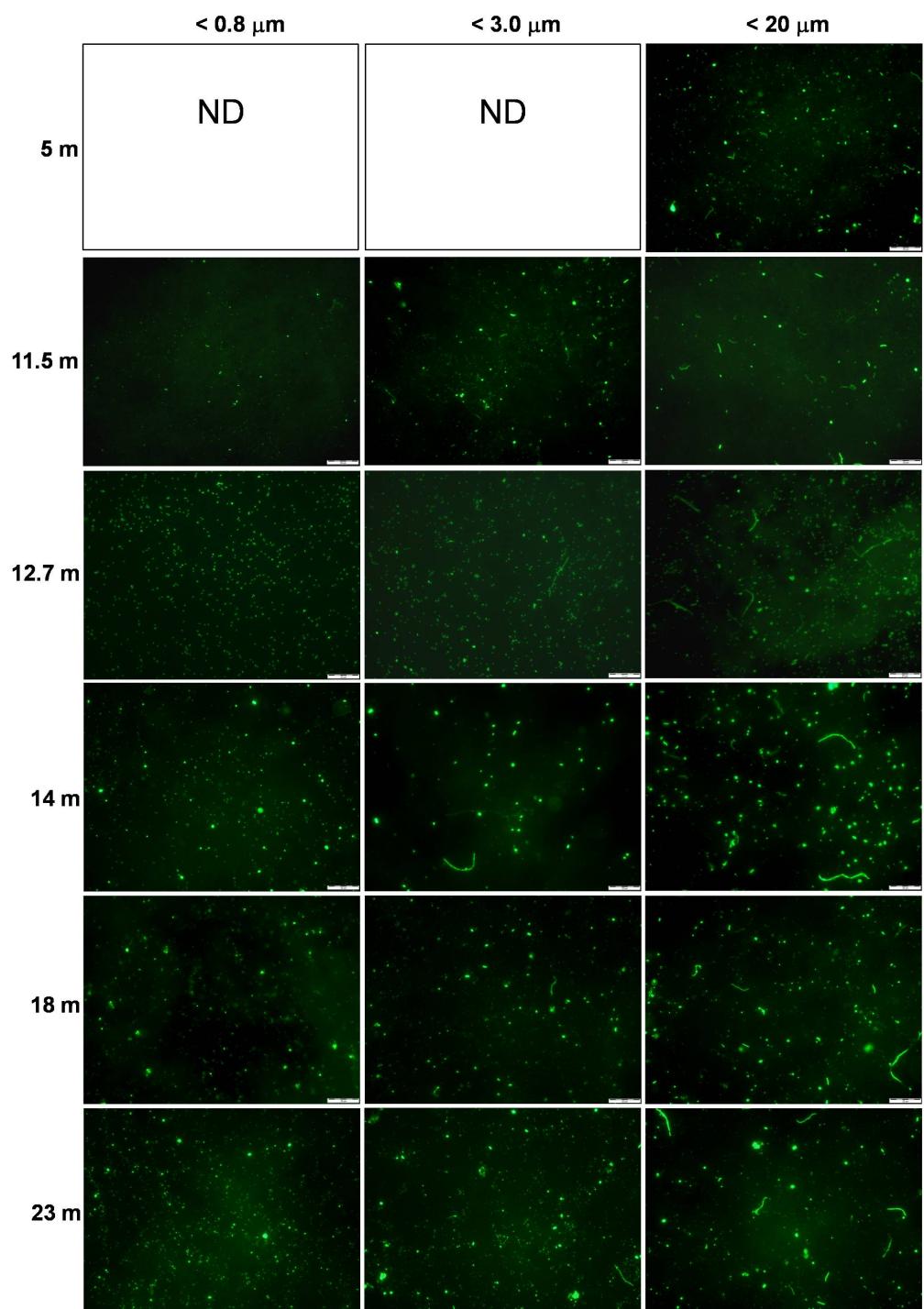


Figure 2.2: Epifluorescence microscopy images of Ace Lake microbiota. Scale bar = 20 μm . ND, not determined.

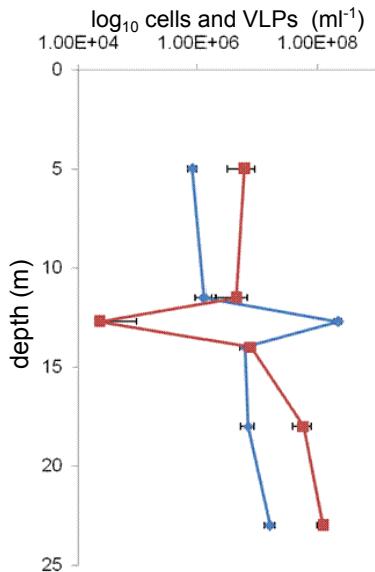


Figure 2.3: Counts of microbial cells (blue) and VLPs (red) by epifluorescence microscopy along a depth profile of Ace Lake. Error bars represent one standard deviation. No VLPs were detected at 12.7 m depth and the value reported represent the detection limit of the counting procedure (*i.e.* one VLP detected in one field of view).

(1.6×10^7 cell ml⁻¹ in the 23 m sample) (Figure 2.3). Cell densities were highest at 12.7 m (2.2×10^8 cells ml⁻¹) corresponding to the GSB layer.

VLP abundance was consistently higher than the cellular counts but the ratio of VLP to cells ranged throughout the water column from ~ 1 –8.5. The exception to this was the 12.7 m where there was an unusual lack of VLPs (Figure 2.3). These data corresponded with the metagenomic data from 12.7 m that found no viral signatures associated with the GSB at the chemo/oxycline (Lauro *et al.*, 2011). From metagenomic data alone the absence of viral signatures in the metagenome does not preclude the presence of viruses with ssDNA or RNA genomes that would not be detected by DNA sequencing method used. Epifluorescence microscopy using SYBR Gold nucleic acid stain would in principle detect VLPs containing non-dsDNA genomes (Patel *et al.*, 2007) that would otherwise be missed in the metagenome. Figure 2.4 contrasts the epifluorescence images of water from 5 m sample with the 12.7 m sample confirming the lack of visible VLPs in the latter. This provided independent support that the GSB population in Ace Lake represents an exception to viral–bacterial population dynamics that describes high rates of genotype cycling in aquatic systems (Rodriguez-Brito *et al.*, 2010).

2.4.3 Development of a metaproteomic mass spectra analysis workflow

The general shotgun proteomics workflow in Figure 2.5 shows that apart from successful sample preparation to simplify the complex protein mixture, protein identification depends greatly upon the post-MS bioinformatic analysis. This is due to how the MS-MS data are used to identify proteins (see Marcotte (2007) for a primer on shotgun

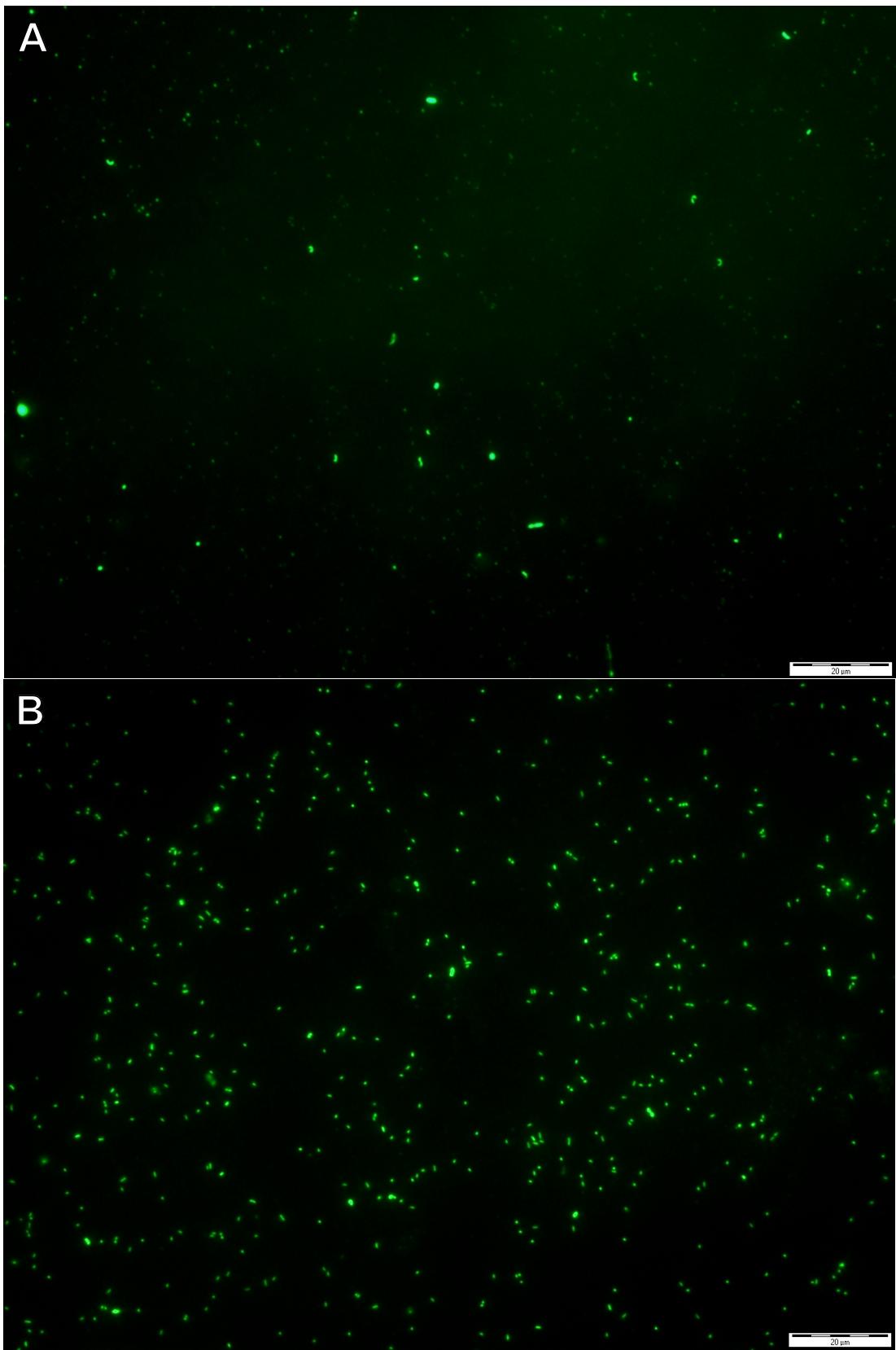


Figure 2.4: Epifluorescence microscopy images contrasting Ace Lake water samples from (A) 5 m and (B) 12.7 m. Numerous VLPs are evident in the 5 m sample, but not in the 12.7 m sample. Scale bar = 20 μ m.

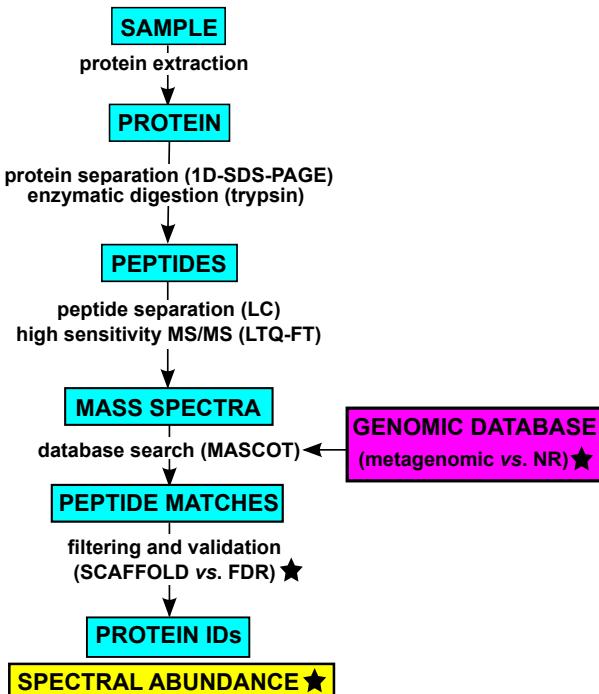


Figure 2.5: A general shotgun proteomics workflow showing how proteins are identified. The procedures or materials used in this study are specified in parentheses. The steps in the workflow that were developed in this study are marked with a star.

proteomic identification).

Briefly, the peptides from digested proteins are separated by LC and subject to a round of mass spectra acquisition where the mass of the peptides (precursor ions) are detected. Selected peptides undergo collision-induced dissociation where they fragment preferentially at the peptide bond. A second round of mass spectra is acquired of the peptide fragments (fragment ion spectra) that represents the amino acid sequence of the peptide. An *in silico* enzymatic digestion is performed on a genomic database to predict the precursor ion masses and their corresponding fragment ion spectra. The fragment ion spectrum and precursor ion mass is used to determine the most likely amino acid sequence of the peptide by comparison to the genomic database and thereby identify the protein(s) of origin. Since spectral matching depends on extremely high mass sensitivity, the ideal genomic database contains complete coverage of sequences from the organism(s) from which the proteins originated. A single amino acid change is sufficient for a peptide match to fail although matching is tolerant to synonymous changes.

Previous metaproteomic analysis was conducted on the 0.1 μm fraction samples along the depth profile of Ace Lake using NR as the search database (Ng, 2010) as metagenomic data were not yet available. The use of a cross-species database requires the genomes to be sufficiently related to identify proteins and application of stringent statistical cut-off to avoid false-positive matches. Across all six samples, a total of 10,443 peptides were identified corresponding to 308 proteins from \sim 400,000 MS-MS spectra (Ng, 2010). Rates of protein identification were low compared to a similar

metaproteomic analysis of the Sargasso Sea, which achieved a total of 5,501 peptide identifications corresponding to 1,042 proteins from ~30,000 MS-MS spectra (Sowell *et al.*, 2009). This indicated that most peptide sequences from Ace Lake were too different from the NR database for protein identification by spectral matching. A key difference in the Sargasso Sea study was the inclusion of metagenomic sequence in the search database from their target populations of SAR11, *Synechococcus*, and *Prochlorococcus* (Sowell *et al.*, 2009). The low identification rate was to be expected as it has been shown that only half the number of proteins are identifiable when using a genomic database that shares 90% amino acid identity with the matching genome (Denef *et al.*, 2007).

Once the metagenomic sequence for the GSB layer became available, re-analysis of the mass spectra to the matched GSB metagenome was performed as this sample had the lowest genomic complexity and was expected to obtain high identification rates (Ng *et al.*, 2010; Ng, 2010). In the re-analysis, 3,970 peptides were identified, mapping to 504 proteins from ~100,000 spectra, which was a near 3-fold increase in the number of protein identifications (Ng *et al.*, 2010; Ng, 2010). This indicated a similar increase in protein identifications could be achieved with the other Ace Lake samples. However, metagenomic sequence from more diverse communities adds additional bioinformatic considerations for protein matching due to their inherently greater heterogeneity. For a protein to be identified according to standard stringency cut-offs it requires at least two peptides to be mapped to it with at least one of those being unique. The converse of this is that it allows for the fact there are potentially non-unique peptides shared by other proteins. Although shared peptides occur between proteins in single genomes, a typical metagenome will contain related species or strains and therefore many more copies of closely related proteins. Protein identification scores in the focussed GSB study was also set to only accept identifications above a determined false discovery rate (FDR). This is the score given by a peptide match to a randomised version of the genomic database (Ng *et al.*, 2010; Ng, 2010).

To address the additional challenges posed by higher diversity metagenomic data, the program SCAFFOLD 2.0 was adopted for the filtering and validation of peptide and protein matches (Figure 2.5). Instead of determining a FDR, SCAFFOLD 2.0 employs the PEPTIDE PROPHET algorithm (Keller *et al.*, 2002) for protein validation. This algorithm fits a distribution of scores from correct and in-correct peptide matches and from this calculates the probability that each result is a genuine match (Keller *et al.*, 2002). More importantly, it identifies groups of proteins that cannot be distinguished based on unique peptides and so accepts all of those proteins in the group may be present in the sample. Two classes of these groups were defined: (1) protein groups, which are proteins with shared peptides that were indistinguishable from the mass spectral analysis and (2) protein ambiguity groups, which are proteins that have some shared peptides. SCAFFOLD 2.0 also facilitates the estimation of protein abundances from spectral counts based on the assumption that proteins that are more abundant will produce more mass spectra. Spectral counting is only semi-quantitative as differences

Table 2.2: Comparison of the number of peptides/proteins identified in the Ace Lake 0.1 μm size fraction metaproteomes using NR vs. matched metagenomic databases. Peptide and protein identifications using NR recorded from Ng (2010). metag, metagenomic database; *a*Proteins identified using SCAFFOLD 2.0; *b*Proteins identified using a FDR.

| Depth (m) | Spectra | Spectra matched (%) (metag) | Peptide IDs (metg) | Protein IDs (metag) | Peptide IDs (NR) | Protein IDs (NR) |
|--------------|---------|--------------------------------------|--------------------------|---------------------------|---------------------|---------------------|
| 5 | 71,201 | 10,843 (15) | 5,728 | 501 | 862 | 15 |
| 11.5 | 53,078 | 6,076 (11) | 3,213 | 224 | 327 | 10 |
| 12.7 | 127,697 | 29,578 (23) | 12,718 | 505a/504b | 4,611 | 169 |
| 14 | 100,650 | 9,008 (9) | 3,427 | 369 | 2,124 | 102 |
| 18 | 131,800 | 1,520 (1) | 725 | 101 | 935 | 11 |
| 23 | 232,797 | 3,648 (2) | 1,602 | 124 | 1,584 | 1 |
| TOTAL | 717,223 | 60,673 (8) | 27,413 | 1,824 | 10,443 | 308 |

between low abundance proteins becomes difficult to gauge. Also, peptides likely have some intrinsic biases in their ionisation and fragmentation efficiencies. For this purpose, defining the protein groups, particularly the ambiguity groups, becomes relevant as it provides a framework to decide how to allocate spectral counts in protein ambiguity groups where spectra are shared. To find statistically significant differences in protein abundances by spectral counting, several normalisation steps had to be incorporated into the metaproteomic analysis workflow and were implemented using in-house scripts developed specifically for this analysis. Differences in the normalised protein abundances from each zone of Ace Lake were then tested using the statistical program STAMP. The steps in the final workflow is detailed in section 2.3 *Materials and methods*.

The final analysis workflow was applied to all the Ace Lake metaproteomic mass spectra datasets. Appendix B lists all proteins identified in Ace Lake using the modified workflow. Identification rates were significantly higher using the matched metagenomic database compared to NR with a 6-fold increase in total proteins identified (Table 2.2). Mass spectra re-analysed from the 12.7 m GSB layer using the modified metaproteomic workflow showed only one additional protein identification (Table 2.2). This indicates that the SCAFFOLD 2.0 protein validation algorithm is as stringent as FDR cut-offs when dealing with lower diversity samples. In addition, 125 of the 505 protein identifications were identified by SCAFFOLD 2.0 to be protein groups (Table B.3). In other words, those mass spectra mapped to two or more proteins making it impossible to distinguish which of those possible proteins were expressed. 11 proteins were also found to be part of an ambiguity group and thus shared peptides with other proteins in the sample (Table B.3). By tracking protein groups, SCAFFOLD 2.0 flags proteins that require an additional validation step, which is to verify if all members of the group have the same functional annotation. All protein groups in the 12.7 m sample that were annotated had identical designations indicating they likely had the same function. The outcomes of the inclusion of spectral counting in the metaproteomic workflow is detailed below.

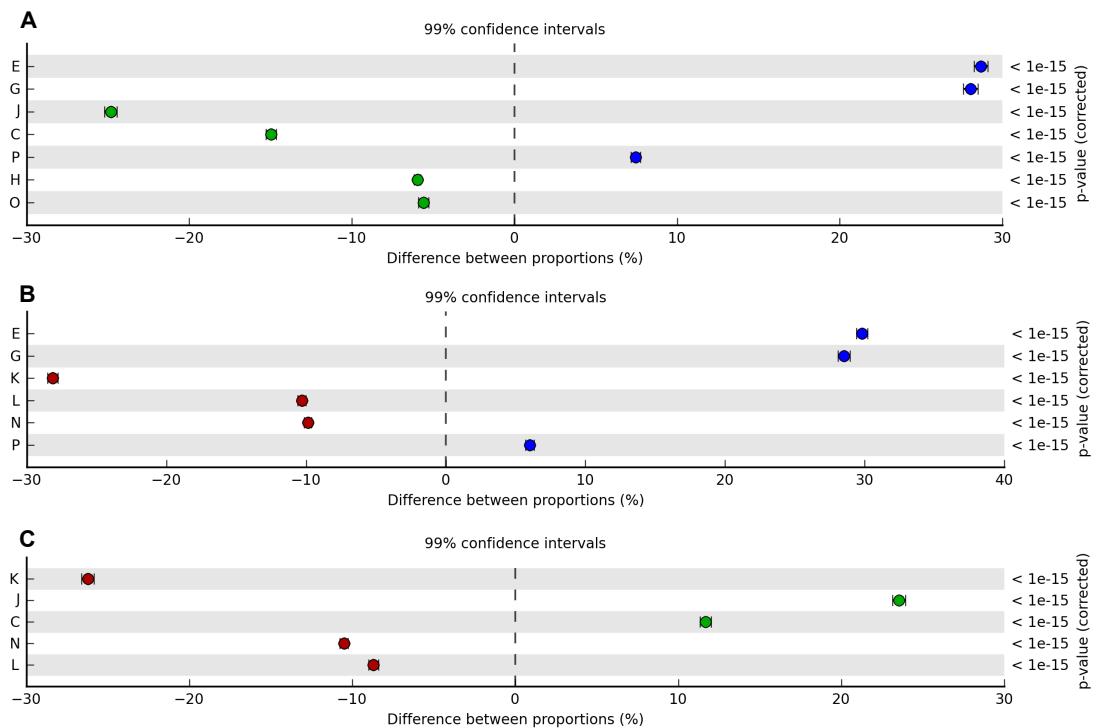


Figure 2.6: Statistical analysis of normalised mass spectra from COG annotated protein between each zone in Ace Lake. Proteins are shown grouped into COG categories. Only categories with corrected p-values <0.05 and effect size >5 are displayed. (A) Mixolimnion compared to chemo/oxycline. (B) Mixolimnion compared to monimolimnion. (C) Chemo/oxycline compared to monimolimnion. Blue, mixolimnion; green, interface; red, monimolimnion. COG category descriptions are: E, amino acid transport and metabolism; G, carbohydrate transport and metabolism; J, translation, ribosomal structure and biogenesis; C, energy production and conservation; P, inorganic ion transport and metabolism; H, co-enzyme transport and metabolism; O, post-translational modification, protein turnover and chaperones; K, transcription; L, replication, recombination and repair; N, cell motility.

2.4.4 Insights from the metaproteomic analysis of Ace Lake

Metaproteomic identifications were able to support biological inferences about each zone in the Ace Lake community. Functions defining each zone was indicated at a broad level by overrepresentation of proteins groups assigned to COG categories (Figure 2.6). This gave an indication of how functional processes were separated with depth. Large numbers of functionally annotated proteins could also be clearly linked to taxonomic groups allowing their contribution to lake ecosystem function and their evolution within the Ace Lake community to be inferred.

Mixolimnion

The most abundant proteins in the mixolimnion were assigned to transport functions from COG categories (E) amino acid transport and metabolism, (G) carbohydrate transport and metabolism and (P) inorganic ion transport and metabolism (Figure 2.6). Most of the protein identifications could be related to taxonomic groups as well as pro-

tein families. The transporters were predominately ATP-binding cassette (ABC) type, with a high COG representation of transporters for carbohydrates (~34% of normalised spectra), amino acids (~32%) and inorganic ions (~9%) (Figure 2.6, Table B.1, Table B.2). The prevalence of amino acid and simple sugar transporters and the low dissolved organic carbon (DOC) concentration in the Ace Lake mixolimnion (Figure 2.1) is likely to reflect efficient utilisation of these substrates from the DOC pool. Thus, examination of the expressed transport proteins may better indicate substrate preferences and nutritional requirements than measurements of nutrient availability.

All transporters in the metaproteome were of bacterial origin and conservative phylogenetic level assignments of the normalised spectral abundance showed the majority to originate from *Proteobacteria* (69%) (of which SAR11 comprised 46%) and *Actinobacteria* 19% (Table B.1, Table B.2). A high proportion of expressed genes with transport functions have also been reported for SAR11 from coastal (Poretsky *et al.*, 2010) and open ocean waters (Sowell *et al.*, 2009; Morris *et al.*, 2010). Oligotrophs, such as SAR11 not only possess a low-diversity of high-affinity transporters (Lauro *et al.*, 2009), but regulate the relative abundance of transporters expressed in response to DOC availability (Poretsky *et al.*, 2010). The transporter expression profile of the Ace Lake SAR11 was very similar to that of the SAR11 in the Sargasso Sea (Sowell *et al.*, 2009). Two SAR11 transport proteins that were detected in Ace Lake (Table B.1, Table B.2) were not detected in the Sargasso Sea (Sowell *et al.*, 2009): an ectoine/hydroxyectoine (167807477 and 167892279) and a zinc ABC transporter (167933120). Ectoine is a compatible solute and presence of the ectoine transporter indicates it is more available in Ace Lake than in the ocean, potentially in response to higher variability in salinity or low temperature. The zinc ABC transporter is likely to support zinc efflux in response to zinc concentrations which are ~70-fold higher in the mixolimnion of Ace Lake compared to seawater (Rankin *et al.*, 1999). Conversely, phosphate transporters were a major class detected from the Sargasso Sea (Sowell *et al.*, 2009) but were absent from the Ace Lake metaproteome consistent with lower phosphate levels in the Sargasso Sea (<5 nM) compared to Ace Lake (1–12 µM). The differences in transporter expression between Ace Lake and oceanic SAR11 are likely to signify adaptive growth strategies that have evolved in the Ace Lake SAR11 community.

Actinobacteria sequences were associated with a diverse phylogenetic cluster (Luna cluster) mainly represented by freshwater ultramicrobacteria (Hahn *et al.*, 2003). Several Luna cluster isolates contain rhodopsin genes, termed actinorhodopsins (Sharma *et al.*, 2009) and similar gene sequences were present in the Ace Lake oxic zone data and found to be expressed (167820670 and 163154474; Table B.1, Table B.2). This was the first report of expression of these actinorhodopsin sequences. The abundance of *Actinobacteria* transporters along with their small cell size and distribution in the water column indicates they occupy a similar ecological niche as SAR11. SAR11 contains proteorhodopsin, which is a related to actinorhodopsin (Sharma *et al.*, 2009). Although the physiological role of proteorhodopsin in SAR11 is yet to be fully elucidated (Fuhrman *et al.*, 2008), this provides some indication actinorhodopsins in Antarctic Luna cluster

Actinobacteria have a related functional role.

Chemo/oxycline

Proteins from the chemo/oxycline were almost all of GSB origin (Table B.3). An in-depth metaproteogenomic analysis of the GSB metabolism has been described (Ng *et al.*, 2010; Ng, 2010). Comparative analysis of proteins between the lake strata showed this depth was more similar to the monimolimnion than the mixolimnion (Figure 2.6). Compared with the mixolimnion, COG categories (J) translation, ribosomal structure and biogenesis; (C) energy production and conservation; (H) co-enzyme transport and metabolism and (O) post-translational modification, protein turnover and chaperones were overrepresented, whereas in comparison with the monimolimnion, only categories J and C were significantly overrepresented. This difference was likely due to the presence of sedimenting GSB cells in the monimolimnion. Nonetheless, overrepresentation of J and C categories indicates it is the GSB population at the chemo/oxycline that is the most metabolically active and productive in the whole lake community. The overrepresentation of category C is similarly evident in the metagenomic comparison of functional genes whereas category J is not (Lauro *et al.*, 2011). This indicates differences in the regulation of energy metabolism compared to protein translation in GSB.

Both metagenomic and microscopic analysis of the GSB layer has indicated the population lacks viruses. Mathematical modeling predicted that the absence of virus predation in the GSB could be an adaptation to longer cycles of growth and inactivity in response to the polar light regime (Lauro *et al.*, 2011). A mechanism for how virus resistance may be conferred in the population was suggested in the metaproteome. Abundant clustered regularly interspaced short palindromic repeat (CRISPR) associated CRISPR-associated proteins (CAS) proteins Cse2, Cse3 and Cse4 (165526330, 165526332 and 165526334, respectively) were detected in the 12.7 m metaproteome (Table B.3). The CAS gene locus (cas3, cse1, cse2, cse3, cse4, cas5, cas1b), to which the proteins map, shares its organisation with CAS loci of sequenced GSB, and groups with the *E. coli* subtype/variant 2. The CRISPR/CAS system has been shown in other organisms to mediate virus resistance (Karginov and Hannon, 2010; Horvath and Barrangou, 2010) and is likely to have a similar function in the Ace Lake GSB.

Monimolimnion

In parallel with taxonomic diversity increasing with depth, with the exception of the GSB layer, the rate of metaproteomic identification of proteins decreased with depth (Table 2.2). Coverage of genomic information and degree of sequence assembly from the more diverse monimolimnion was not as high as for the mixolimnion or the chemo/oxycline. This means that more peptides in the monimolimnion than in the other depths mapped to fragmentary or absent metagenomic data and failed to be identified. Annotated proteins in the monimolimnion were overrepresented in COG categories (K) transcription; (L) replication, recombination and repair and (N) motility

(Figure 2.6). Categories L and N were similarly overrepresented in the monimolimnion metagenomic samples (Lauro *et al.*, 2011) demonstrating the genomic expansion of these functions correlated with higher abundance of these proteins. However, differential abundance was greatest in the category K proteins, which showed little difference in relative abundance in the metagenomes (Lauro *et al.*, 2011) suggesting transcription proteins in the monimolimnion were up-regulated (Figure 2.6). The majority of the proteins that were detected in the monimolimnion (e.g. 67% at 23 m) (Table B.6) were for hypothetical proteins that tended to lack orthologues in well-characterized organisms. This demonstrates the extremely high level of functional novelty in the anaerobic zone of the lake. These hypothetical proteins represent potential targets for protein expression studies to determine their functional properties.

2.5 Conclusions

In this study an epifluorescence microscopy and a metaproteomic analysis workflow were developed that were able to address two key areas that metagenomics sequencing alone does not. The first of these was determining the density of cells and VLPs and assessment of gross morphology. The counts and microscopy images validated the size fractionation procedure, showed the stratification of the lake populations with depth and suggested the unusual absence of VLPs in association with the GSB layer. As a lack of viral signatures was observed by both microscopy and metagenomic analysis in the GSB layer, this provided independent lines of support that viruses were indeed absent or rare. The second was identifying which proteins were expressed by the microbial community. The bioinformatic analysis workflow specifically tailored to analyse metaproteomic mass spectral data afforded an increase in protein identifications, identified proteins with shared peptides and enabled estimation of protein abundances by spectral counting. This identified differences in the functional complement between the three strata of the lake. The mixolimnion was shown to be dominated by transporter proteins; the chemo/oxycline by energy production and biosynthetic functions; and the monimolimnion by proteins involved in transcription, replication, recombination and motility. In the lower complexity mixolimnion and chemo/oxycline samples better designations of protein functions and clear links to their taxonomic origins were possible. These provided insights into substrate usage of the dominant aerobic bacterial populations and specific adaptations to the Antarctic lake environment compared to the ocean and suggested mechanism for virus resistance in the GSB. Both the microscopic and metaproteomic analyses have added crucial information to the metagenomic datasets allowing for an integrative understanding of the whole lake ecosystem. Furthermore, these have been applied to subsequent studies of Organic Lake described in chapters 3 and 4.

Chapter 3

Virophage control of Antarctic algal host–virus dynamics

Co-authorship statement

A version of this chapter has been published as:

Sheree Yau, Federico M. Lauro, Matthew Z. DeMaere, Mark V. Brown, Torsten Thomas, Mark J. Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffrey M. Hoffman, John A. Gibson, and Ricardo Cavicchioli. Virophage control of antarctic algal host–virus dynamics.

Proceedings of the National Academy of Sciences USA 108:6163–6168, 2011.

Contributions to this publication by other researchers is as follows. Research was designed and the manuscript edited by Federico Lauro, Mark Brown, Torsten Thomas, John Gibson and Ricardo Cavicchioli. Sample collection was performed by Federico Lauro, Mark Brown, Torsten Thomas, Jeffrey Hoffman and Ricardo Cavicchioli. DNA extraction and clone library preparation of 2006 samples was performed by Cynthia Andrews-Pfannkoch and Jeffery Hoffman. DNA sequencing quality control was performed by Matthew Lewis. Metagenomic sequence filtering, global assembly and annotation was performed by Matthew DeMaere. Assistance in mass spectrometry was provided by Mark Raftery. Assistance in analysis of eucaryal taxonomy was provided by Mark Brown. Analysis of virophage abundance over time was performed by Federico Lauro.

Apart from these contributions, I performed all other data analyses, interpretations and drafted the manuscript.

3.1 Abstract

Viruses are abundant ubiquitous members of microbial communities, and in the marine environment affect population structure and nutrient cycling by infecting and lysing primary producers. Antarctic lakes are microbially dominated ecosystems supporting truncated food webs where viruses exert a major influence on the microbial loop. Here we report the discovery of a new virophage (relative of the recently described Sputnik virophage) that preys on phycodnaviruses that infect prasinophytes (phototrophic algae). By performing metaproteogenomic analysis on samples from Organic Lake, a hypersaline meromictic lake in Antarctica, complete virophage and near-complete phycodnavirus genomes were obtained. By introducing the virophage as an additional predator of a predator-prey dynamic model we determine that the virophage stimulates secondary production through the microbial loop by reducing overall mortality of the host and increasing the frequency of blooms during polar summer light periods. Virophages remained abundant in the lake two years later, and were represented by populations with a high level of major capsid protein sequence variation (25–100% identity). Virophage signatures were also found in neighbouring Ace Lake (in abundance), and in two tropical lakes (hypersaline and fresh), an estuary, and an ocean upwelling site. These findings indicate that virophages regulate host–virus interactions and influence overall carbon flux in Organic Lake, and play previously unrecognised roles in diverse aquatic ecosystems.

3.2 Introduction

It has been known for at least 20 years that viruses frequently infect and lyse marine primary producers causing up to 70% of cyanobacterial mortality (Proctor and Fuhrman, 1990; Suttle *et al.*, 1990). Eucaryotic phytoplankton are preyed upon by large dsDNA phycodnaviruses (PVs) causing bloom termination in globally distributed species (Nagasaki *et al.*, 1994; Jacobsen *et al.*, 1996; Wilson *et al.*, 2002; Martínez-Martínez *et al.*, 2007). Elevated levels of dissolved organic carbon (DOC) (Eberlein *et al.*, 1985) and numbers of heterotrophic bacteria (Davidson and Marchant, 1992; Bratbak *et al.*, 1998; Castberg *et al.*, 2001) occur during algal blooms indicating that viral lysis of eucaryotic algae stimulates secondary production. Viruses also suppress host populations at concentrations below bloom-forming levels, with abundance being controlled by the efficiency and production rates of the infecting viruses (Larsen *et al.*, 2001; Bouvier and del Giorgio, 2007). Antarctic lakes are microbially dominated ecosystems supporting few, if any metazoans in the water column (Laybourn-Parry, 1997). In these truncated food webs, viruses are expected to play an increased role in the microbial loop (Madan *et al.*, 2005). Low complexity Antarctic lake systems are amenable to whole community based molecular analyses where the role that viruses play in microbial dynamics can be unravelled (Lauro *et al.*, 2011). Attesting to this, a metagenomic study of Lake Limnopolar, West Antarctica uncovered a dominance of eucaryotic viruses and ssDNA viruses previously unknown in aquatic systems (López-Bueno *et al.*, 2009).

Organic Lake is a shallow (~7 m) hypersaline (~230 g L⁻¹ maximum salinity) meromictic lake in the Vestfold Hills with a high concentration of dimethylsulphide (DMS) (~120 µg L⁻¹) in its bottom waters (Gibson *et al.*, 1991; Roberts *et al.*, 1993). Water temperature at the surface of the lake can vary from -14 to +15°C while remaining sub-zero at depth (Franzmann *et al.*, 1987b; Gibson, 1999). The lake is eutrophic, with organic material sourced both from autochthonous production and potential input from penguins and terrestrial algae. The high concentrations of organic material reflect slow breakdown in the highly saline lake water. The salt in the lake was trapped along with the marine biota when the lake was formed due to falling sea level ~3,000 BP (Bird *et al.*, 1991; Zwart *et al.*, 1998). Low species diversity (Shannon-Weaver diversity: 1.01) and richness (Chao non-parametric index: 32 ± 12) has been reported for the sediment community (Bowman *et al.*, 2000a). Unlike high latitude lakes, viral abundance has been reported to increase with trophic status (Madan *et al.*, 2005) and with salinity in Antarctic lakes (Laybourn-Parry *et al.*, 2001).

In order to functionally characterise its microbial community, a metaproteogenomic program for Organic Lake was established. This study reports the analysis of the surface water of Organic Lake, highlighting the presence of a relative of the recently described Sputnik virophage, a small eucaryotic virus that requires a helper *Acanthamoeba polyphaga* mimivirus (ApMV) to replicate (La Scola *et al.*, 2008). From metagenomic DNA, a complete Organic Lake virophage (OLV) genome was constructed (the second virophage genome to be described), and near-complete genomes of its probable helper Organic Lake phycodnavirus (OLPV).

3.3 Materials and methods

3.3.1 Samples and DNA sequencing

Water samples collected from Organic Lake were: 1) Surface water from the eastern side of the ice-free lake ($68^{\circ}27'25.48''S$, $78^{\circ}11'28.06''E$) December 24, 2006. 2) A depth profile collected through a 30 cm hole drilled through the surface ice above the deepest point in the lake ($68^{\circ}27'22.15''S$, $78^{\circ}11'23.95''E$), November 10, 2008. 3) Surface water from the north-east side of the partially ice-covered lake ($68^{\circ}27'21.02''S$, $78^{\circ}11'42.42''E$), December 12, 2008. Samples were sequentially filtered through a 20 μm pre-filter and biomass captured onto 3.0, 0.8 and 0.1 μm membrane filters as described previously (Ng *et al.*, 2010; Lauro *et al.*, 2011). The samples from 2008 also included 50% (v/v) RNAlater. DNA extraction, sequencing and quality validation was performed as previously described (Ng *et al.*, 2010; Lauro *et al.*, 2011). DNA sequencing was performed at the J. Craig Venter Institute in Rockville, MD, USA.

3.3.2 Transmission electron microscopy

Unfiltered Organic Lake surface water from December 24, 2006 (fixed on-site in 1% (v/v) formalin) was concentrated and a solvent exchange performed with sterile filtered ammonium acetate solution 1% (w/v) using a 50 kDa cut-off Microcon centrifugal filter device (Millipore) according to the manufacturers instructions. Formvar coated 200 mesh copper grids were floated on a droplet of sample for 30 min, excess liquid wicked off and the grid negatively stained for 30 s with uranyl acetate 2% (w/v). The sample was visualised using a JEOL1400 transmission electron microscope at 100 kV at 150,000 to 250,000 \times magnification.

3.3.3 Metagenomic assembly and annotation

Mosaic metagenomic assemblies were generated as previously described (Ng *et al.*, 2010; Lauro *et al.*, 2011). For the 0.1 μm Organic Lake 2006 sample, assembly was a hybrid of Sanger and 454 read data (Table 3.2). For all other sample size fractions, runtime parameters used were standard for 454 sequencing data. Low GC ($\geq 51\%$) scaffolds >10 kbp from the 0.1 μm 2006 assembly had high coverage ($>45\times$) indicating these were from the dominant taxa. One of these scaffolds was binned as virophage and the rest as PV.

To further separate the OLPV types and assess the completeness of their genomic content, highly conserved single copy PV orthologues were identified as follows. An all against all basic local alignment search tool (BLAST) search was conducted with protein sequences from the ten available PV genomes (*Acanthocystis turfacea* chlorella virus 1, PbCV-1, PbCV AR158, PbCV FR483, PbCV NY2A, *Emiliania huxleyi* virus 86, *Ectocarpus siliculosus* virus 1, *Feldmannia* sp. virus, *Ostreococcus* virus 5, *Ostreococcus tauri* virus 1 and ApMV (which was included as a close PV relative). BLASTp results were parsed and clustered using ORTHOMCL V1.4 (Li *et al.*, 2003; Chen *et al.*, 2006).

Table 3.1: List of primers used to close the OLV genome. Dir, direction.

| Primer function | ID | Dir. | Sequence (5'-3') | Length (bp) |
|---------------------|------|---------|--|-------------|
| Outer gap spanning | SY11 | forward | TTG TCT TAT GTA TTA CAA ATC ATT GAA | 3,843 |
| Outer gap spanning | SY12 | reverse | CGA CAT TAA TCG GTT GTT TT | |
| Nested gap spanning | SY13 | forward | GCA TTA CGA ATG TGT TCC AG | 3,403 |
| Nested gap spanning | SY14 | reverse | TTC TCC GTG ATT GAT ATC GT | |
| Sequencing | SY23 | forward | TCC CTA TTG ATG TCA AAA CC | - |
| Sequencing | SY24 | forward | GAT TCT GGT TGG AGC ATA TAT TT | |

Pairs of each orthologue were located on eight of the PV scaffolds. The location of each orthologue pair had a complementary distribution so the eight scaffolds were able to be sorted unambiguously into two strains (OLPV-1 and OLPV-2). OLPV-1 ribonucleotide reductase α -subunit appeared as duplicated on different scaffold ends, likely as an artefact of its proximity to an assembly break point. The remaining high coverage scaffolds were searched for predicted proteins present in one OLPV strain but not in the other and assigned to the strain in which it was absent. Comparison of OLPV-1 and OLPV-2 scaffolds was performed using tBLASTN of concatenated scaffolds from each strain and visualised using the Artemis comparison tool (ACT) (Carver *et al.*, 2005). DNA sequence data is available in Genbank and Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (<http://web.camera.calit2.net>).

3.3.4 Organic Lake virophage genome completion and annotation

The high coverage ($77\times$), large number of Sputnik homologues that encode essential functions and length of the putative OLV scaffold from the $0.1\text{ }\mu\text{m}$ 2006 hybrid assembly indicated it was a near-complete genome. Reads from this scaffold were reassembled at high stringency and visualised using PHRED/PHRAP/CONSED (Gordon, 2004) to complete the sequence. Mate-pair data indicated a circular molecule and primers were designed to span the ends of the scaffold and sequence across the gap (Table 3.1).

Touch-down polymerase chain reaction (PCR) was performed with DNA from $0.1\text{ }\mu\text{m}$ 2006 sample, the product used for nested PCR and the final product was cloned and sequenced. The complete genome was manually annotated and visualised using ARTEMIS (Rutherford *et al.*, 2000). Translated open-reading frames (ORFs) (minimal size 120 amino acids) were compared (BLASTP) to GenBank, to the all metagenomic ORF peptide database on CAMERA (<http://web.camera.calit2.net>) and to predicted proteins from OLPV-1 and OLPV-2 scaffolds. Comparisons between the OLV genome and OLPV-1/OLPV-2 were performed with tBLASTN and visualised using ACT (Carver *et al.*, 2005).

3.3.5 Phylogenetic analysis

Translated amino acid sequences from viral marker genes of interest were retrieved from the 0.1 μm 2006 metagenomic assemblies from this study, GenBank and CAMERA all metagenomic reads ORF peptide database. Homologous sequences were aligned using MUSCLE v3.6 (Edgar, 2004). Neighbour-joining analysis, test for clade support (bootstrap analysis 2000 replicates) and tree drawing was performed with Molecular Evolutionary Genetic Analysis (MEGA) software v4 (Kumar *et al.*, 2008). Maximum likelihood analysis (JTT substitution model) and test for clade support (aLRT analysis) was performed with PHYML (10) and the tree visualised using MEGA. 18S ribosomal RNA (rRNA) gene sequences were retrieved from reads of all filter sizes, compared (BLASTN, e-value $<1.0\text{e}{-}5$) to the SILVA100 SSURef database, aligned and phylogeny performed using ARB as previously described (Ng *et al.*, 2010; Lauro *et al.*, 2011). The abundance and similarity of virophages in all lake samples and filter sizes was estimated using BLASTP (evalue $<1.0\text{e}{-}5$) to search using the OLV major capsid protein (MCP) sequence against a database of proteins predicted from sequencing reads. The database was generated as previously described (Proctor and Fuhrman, 1990) and the percent identity of the BLAST hit was used as a proxy for species similarity.

3.3.6 Metaproteomic analysis

Extraction of proteins from the 0.1 μm filter from 2006, one dimensional-sodium dodecyl sulphate polyacrylamide gel electrophoresis (1D-SDS PAGE), liquid chromatography (LC) and two dimensional mass spectrometry (MS-MS) was performed as previously described (Ng *et al.*, 2010; Lauro *et al.*, 2011). Spectral matching of MS-MS data was performed as described in chapter 2, *Materials and methods* 2.3 using metagenomic sequences from Organic Lake as the search database. The protein sequence database was generated by combining ORFs from the 3.0, 0.8 and 0.1 μm mosaic assemblies, which comprised 130,581 sequences. SCAFFOLD 3.0 (Proteome Software Inc.) was used to validate MS/MS based peptide and protein identifications using the same stringency parameters as described in chapter 2, *Materials and methods* 2.3). The peptide sequences by which the proteins were identified shown in Appendix Table C.1.

3.3.7 Model of algal host–virus and virophage dynamics

To model the effect a virophage would have on algal *Pyramimonas* algal host populations in Organic Lake, modified Lotka-Volterra equations were used describing the OLV as a predator of predator OLPV. The original equations are given by:

$$\frac{dA}{dt} = \alpha A - \varepsilon PA \quad (3.1)$$

$$\frac{dP}{dt} = \theta PA - \mu P \quad (3.2)$$

Where:

A is the number of *Pyramimonas* (prey).

P is the number of OLPV (predator).

α is the specific growth rate of the prey.

θ is the specific production rate of the predator.

ε is the rate of predator mediated death of prey.

μ is the specific decay rate of the predator.

Equation 3.1 describes the change in *Pyramimonas* abundance and equation 3.2 the change in OLPV abundance in the absence of OLV. In the presence of OLV, *Pyramimonas*, OLPV and OLV dynamics are described by the following equations:

$$\frac{dP}{dt} = \theta PA - \mu P - \omega PV \quad (3.3)$$

$$\frac{dV}{dt} = \beta PV - \gamma V \quad (3.4)$$

Where:

V is the number of the OLV (predator of predator).

ω is the rate of OLV mediated reduction in OLPV infective particles.

β is the production rate of OLV.

γ is the decay rate of OLV.

Equation 3.3 is a modified version of equation 3.2 which includes the effect of OLV on the change in abundance of OLPV. Equation 3.4 describes the growth properties of OLV as a predator of OLPV. Values for the variables for the solution shown (Figure 3.9) were as follows: initial prey (10), predator (1) and predator of predator (10) numbers, $\alpha = 0.1$, $\theta = 0.0015$, $\varepsilon = 0.01$, $\mu = 0.01$, $\omega = 0.01$, $\beta = 0.015$ and $\gamma = 0.15$. COMplex Pathway Simulator (COPASI) software (Hoops *et al.*, 2006) was used to simulate prey, predator and predator of predator dynamics using the deterministic (LSODA) method.

3.4 Results and discussion

3.4.1 Dominance of phycodnaviruses in Organic Lake

Water samples from Organic Lake were collected December 2006 and November and December 2008, and microbial biomass collected onto 3.0, 0.8 and 0.1 μm membrane filters as described previously (Lauro *et al.*, 2011). A large proportion of shotgun sequencing reads (96.2%) from the 0.1 μm size fraction of the 2006 Organic Lake metagenome (Table 3.2) had no significant hits to sequences in the RefSeq database (TBLASTx with

Table 3.2: Summary of metagenomic data for Organic Lake 0.1 μm fraction samples used in this study. SCF, scaffolds.

| ID | Date | Trimmed reads | SCFs >10 kbp (reads in SCFs) | Annotated ORFs in SCFs (total ORFs) |
|-------|---------------|---|---------------------------------|--|
| GS233 | December 2006 | 418,265 (Sanger: 28,481) (454: 389,784) | 45 (221,573) | 7,318 (21,961) |
| GS374 | November 2008 | 454: 494,573 | 5 (771) | 33,262 (83,684) |
| GS379 | December 2008 | 454: 446,200 | 2 (40,314) | 23,012 (64,779) |

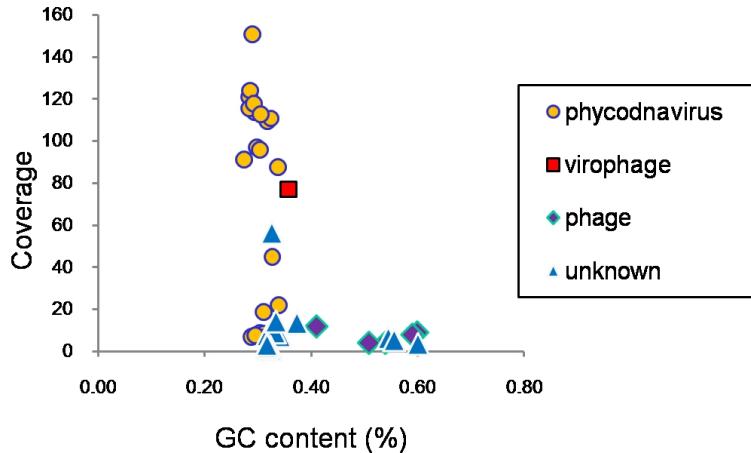


Figure 3.1: Plot of percent GC content *vs* coverage for the 2006 Organic Lake 0.1 μm hybrid assembly scaffolds >10 kb.

e-value $<1.0\text{e}-3$, minimum alignment length: 60 bp, minimum identity: 60%). The degree of assembly was high, with 77% of reads forming part of a scaffold, indicating the sample contained a few abundant taxa of minimal diversity. Forty-five scaffolds were longer than 10 kbp; the five longest ranged from 70 to 171 kbp. GC content and coverage were used to separate scaffolds into taxonomic groups (Figure 3.1). A broad division was evident between low ($\leq 41\%$) and high ($\geq 51\%$) GC scaffolds suggesting they constituted two taxonomic groups. All scaffolds in the high GC group that could be assigned contained phage homologues, as did the one exceptional low GC scaffold. The low coverage in the high GC group showed bacteriophages were not abundant in the 0.1 μm fraction. These scaffolds were not analysed further. The low GC scaffolds with confident assignments contained sequences matching conserved phycodnavirus (PV) or *Acanthamoeba polyphaga* mimivirus (ApMV) proteins. These PV-related scaffolds comprised 60% of assembled reads demonstrating that Organic Lake phycodnaviruses (OLPVs) were numerically dominant in the 0.1 μm fraction. transmission electron microscopy (TEM) revealed the presence of virus-like particles with the dimensions and structure typical of PVs (Figure 3.2A).

Within the low GC group, scaffolds separated into a high coverage ($>45\times$) group, including the five longest scaffolds, and a low coverage ($<22\times$) group. Two of the scaffolds in the high coverage group and one in the low coverage group contained the PV marker DNA polymerase B (DPOB). The two high coverage DPOB share 76%

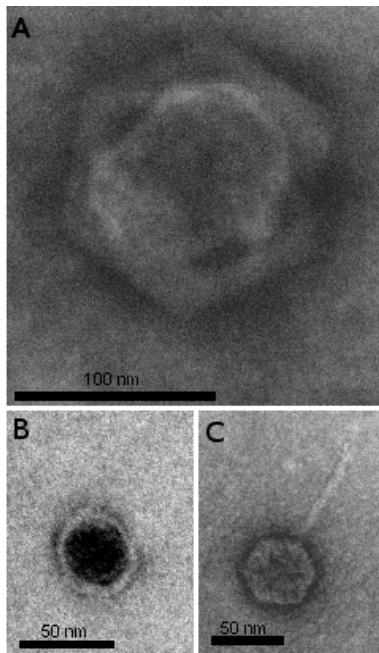


Figure 3.2: Transmission electron micrographs of negatively stained VLPs from Organic Lake. (A) VLP resembling the size and morphology of PVs, (B) Sputnik virophage and (C) bacteriophages.

amino acid identity and both share ~57% identity to the low coverage DPOB. DPOB is single-copy throughout the nucleo-cytoplasmic large DNA virus (NCLDV) family to which PVs belong (Iyer *et al.*, 2001, 2006), demonstrating that the Organic Lake surface waters contained two closely related abundant PV types (DPOB1) and (DPOB2), and a more distantly related lower abundance type (DPOB3).

Phylogenetic analysis clustered Organic Lake DPOB with unclassified lytic marine PV isolates that infect the prymnesiophytes *Chrysochromulina ericina* (CeV1) and *Phaeocystis pouchetii* (PpV), the prasinophyte *Pyramimonas orientalis* (PoV) (Jacobsen *et al.*, 1996; Sandaa *et al.*, 2001), and uncultured marine PVs related to ApMV (Monier *et al.*, 2008b,a) (Figure 3.3). As the host range of PVs broadly correlates with DPOB phylogeny (Nagasaki *et al.*, 2005; Larsen *et al.*, 2008), OLPVs would infect prasinophytes or prymnesiophytes. The most probable host is the prasinophyte, *Pyramimonas* as no prymnesiophyte 18S ribosomal RNA (rRNA) gene sequences were present in any size fraction of the Organic Lake metagenome (Figure 3.4). Neither were prymnesiophytes found in subsequent analysis of the Organic Lake profile in November 2008 (see chapter 4) suggesting they are rare or absent in the Organic Lake community.

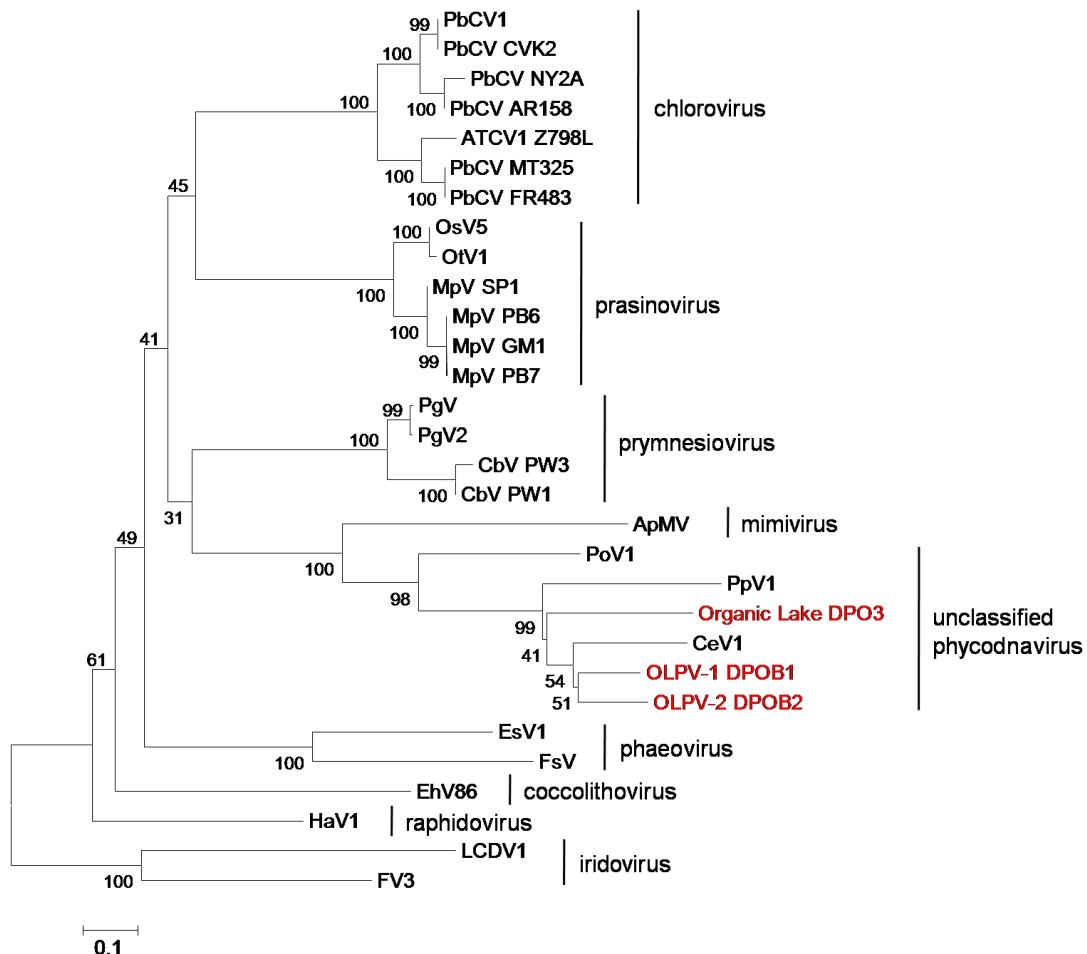


Figure 3.3: Neighbour-joining tree of B family DNA polymerase amino acid sequences from OLPV and NCLDV sequences from GenBank. Organic Lake sequences are shown in red. Abbreviations and accession numbers from bottom to top: PbCV1, *Paramecium bursaria* chlorella virus 1 (AAC00532.1); PbCV CVK2, *P. bursaria* chlorella virus CVK2 (BAA35142.1); PbCV NY2A, *P. bursaria* chlorella virus NY2A (ABT14648.1); PbCV AR158, *P. bursaria* chlorella virus AR158 (ABU43776.1); AtCV1, *Acathocystis turfacea* chlorella virus (ABT16932.1); PbCV MT325, *P. bursaria* chlorella virus MT325(ABT13573.1); PbCV FR483, *P. bursaria* chlorella virus FR483 (ABT15308.1); OsV5, *Ostreococcus* virus 5 (ABY28020.1); OtV1, *O. tauri* virus 1(YP_003495047.1); MpV SP1, *Micromonas pusilla* virus SP1(AAB66713.1); MpV PB6, *M. pusilla* virus PB6 (AAB49743.1); MpV GM1, *M. pusilla* virus GM1 (AAB49742.1); MpV PB7, *M. pusilla* virus PB7 (AAB49744.1); CbV PW1, *Chrysochromulina brevifilum* virus PW1 (AAB49739.1); CbV PW3, *C. brevifilum* virus PW3 (AAB49740.1); ApMV, *Acathamoeba polyphaga* mimivirus (AAV50591.1); PoV, *Pyramimonas orientalis* virus (ABU23717.1); PpV, *Phaeocystis pouchetii* virus (ABU23718.1); CeV1, *C. ericinia* virus 1 (ABU23716.1); EsV1, *Ectocarpus siliculosus* virus (AAK14511.1); FsV, *Feldmannia* sp. virus (AAB67116.1); EhV86, *Emiliania huxleyi* virus 86 (CAI65453.1); HaV1, *Heterosigma akashiwo* virus 1 (BAE06251.1); FV3, Frog virus 3(AAT09720.1) and LCDV1, Lymphocystis disease virus 1(NP_078724.1).

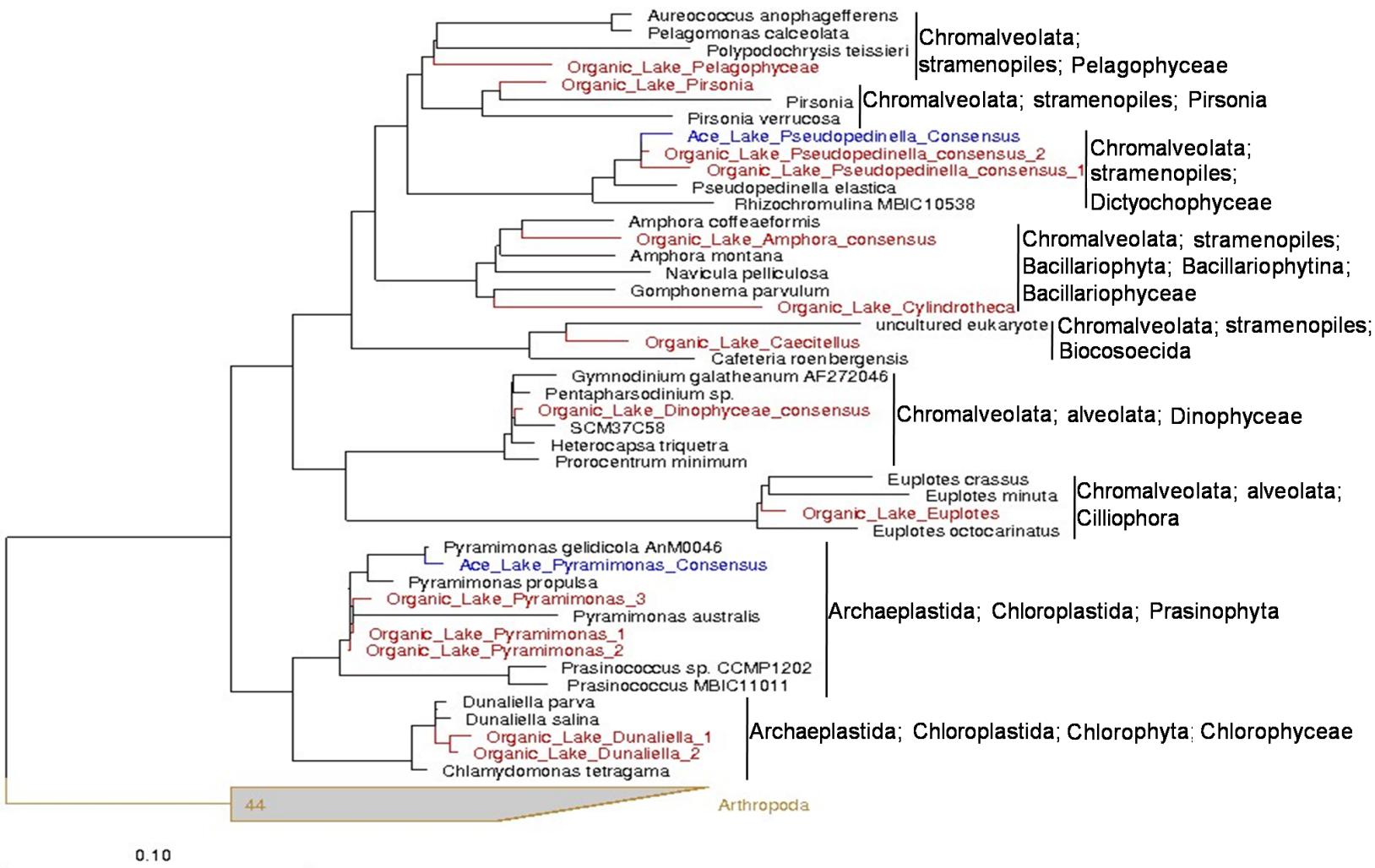


Figure 3.4: Phylogeny of the 18S rRNA genes from the 2006 Organic Lake metagenome. Organic Lake sequences are shown in red. Ace Lake sequences are shown in blue.

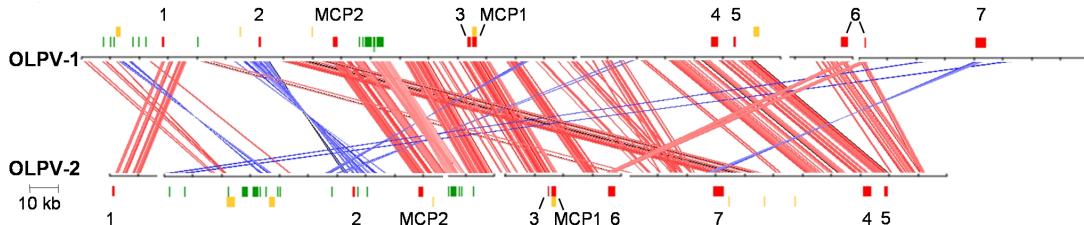


Figure 3.5: Maps of OLPV-1 and OLPV-2 scaffolds and comparison of the location of genes. Genes are marked as follows: single-copy conserved orthologues and MCP (red), regions with identity to OLV (green), proteins identified in the metaproteome (yellow), ribosomal nucleotide reductase β (1), VV A32 packaging ATPase (2), VV VLTF3 transcription factor (3), VV D5 replicative helicase (4), PbCV-1 A482R-like putative transcription factor (5), ribonucleotide reductase α (6), and DNA polymerase B (7). Lines connect homologous regions between the OLPV-1 and OLPV-2 scaffolds in the same orientation (red) and reverse orientation (blue).

Supporting the presence of more than one PV, pairs of single-copy PV orthologues (ribonucleotide reductase α and β subunits, VV A32R virion packaging helicase, PBCV1 A482R-like putative transcription factor, VV D5 ATPase and VLTF2 family transcription factor) were identified in the high coverage scaffolds that shared an average of 81% percent amino acid identity. Based on the positions of single copy genes on the scaffolds and the percent identity between them, the high coverage scaffolds were grouped into two strains designated OLPV-1 and OLPV-2 according to their DPOB phylogeny (Figure 3.3). The remaining high coverage scaffolds were assigned to either strain, resulting in two near-complete genomes of \sim 300 kbp each (Figure 3.5), that are within the range of other sequenced PV genomes (155–407 kbp). In addition, several OLPV genomic fragments contained PV homologues in high coverage scaffolds that could not be confidently assigned to either strain.

Both OLPV strains contain a PpV-like major capsid protein (MCP) designated MCP1 and another unique MCP designated MCP2 (Figure 3.6). Both OLPV MCP1s were identified in the metaproteome (Figure 3.5 and Table 3.3) but MCP2 was not. In addition to MCPs, the metaproteome contained a range of abundant structural proteins and others more likely to be packaged in the virion (e.g. chaperone), that were expressed by OLPV-1, OLPV-2 and/or an OLPV genomic fragment Table 3.3. These data suggest that MCP1 is the major structural protein, and that both OLPV-1 and OLPV-2 were in a productive cycle in the lake at the time of sampling.

3.4.2 Complete genome of an Organic Lake virophage

Sputnik is a small (50 nm) icosahedral satellite virus of mamavirus (a new strain of ApMV). It was termed a virophage because co-infection with Sputnik is deleterious to the mamavirus, resulting in abnormal virions and a decrease in mamavirus infectivity (La Scola *et al.*, 2008). One 28 kbp scaffold in the low GC high coverage group had six out of 38 predicted proteins homologous to Sputnik virophage proteins (Figure 3.7 and Table 3.4), and one PV homologue. The scaffold had a low GC content (\sim 30%), similar

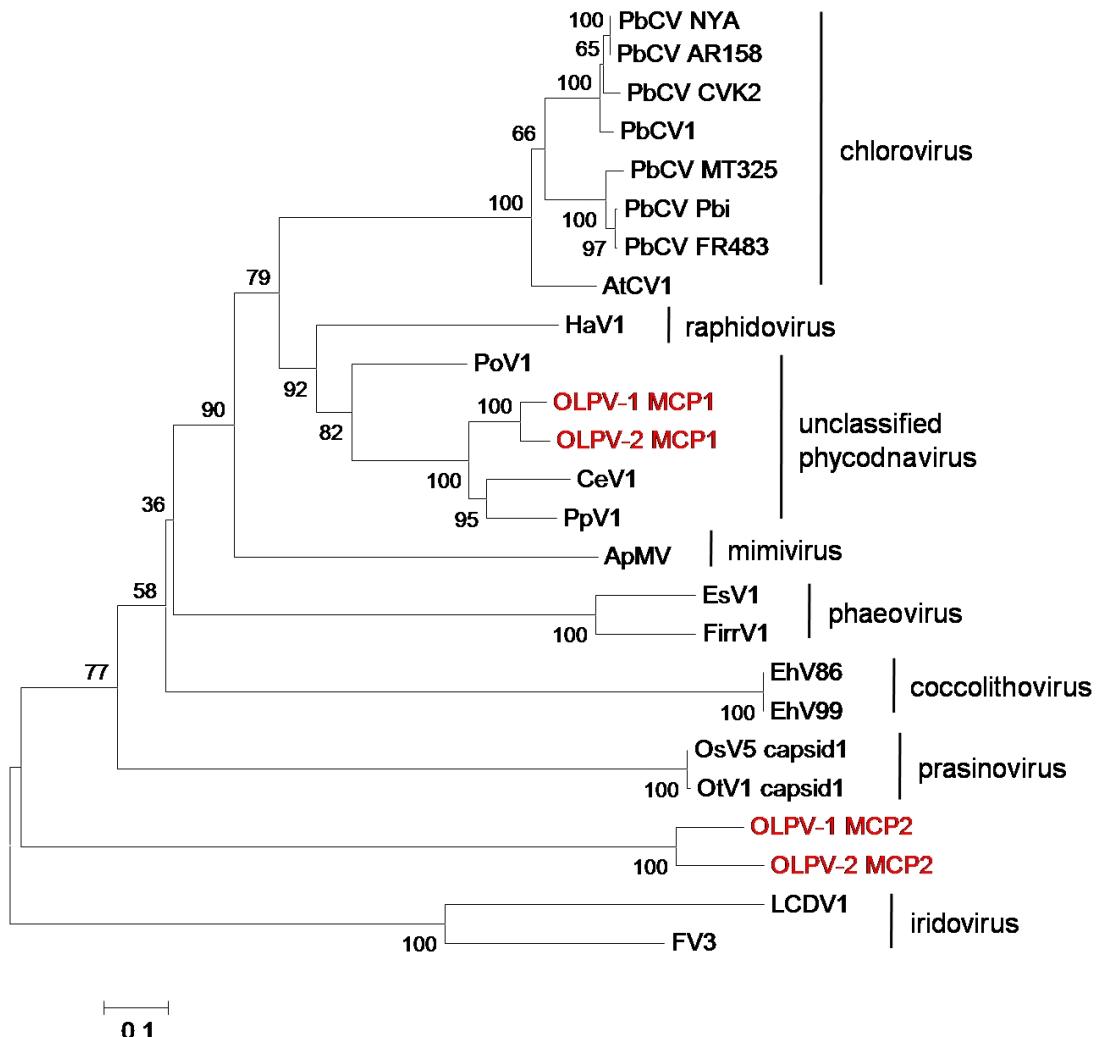


Figure 3.6: Neighbour-joining tree of major capsid protein amino acid sequences from OLPV and other NCLDV sequences from GenBank. Abbreviations and accession numbers from top to bottom: PbCV NYA, *Paramecium bursaria* chlorella virus NY2A (ABT14984.1); PbCV AR158, *P. bursaria* chlorella virus AR158 (ABU44077.1); PbCV CVK2, *P. bursaria* chlorella virus CVK2 (BAA35143.1); PbCV1, *P. bursaria* chlorella virus 1(AAA88828.1); PbCV MT325, *P. bursaria* chlorella virus MT325 (ABT14017.1); PbCV Pbi, *P. bursaria* chlorella virus Pbi (AAC27492.1); PbCV FR483, *P. bursaria* chlorella virus FR483 (ABT15755.1); AtCV1, *Acathocystis turfacea* chlorella virus 1 (ABT16414.1); HaV1, *Heterosigma akashiwo* virus 1 (BAE06835.1); PoV, *Pyramimonas orientalis* virus (ABU23714.1); CeV1, *Chrysochromulina ericinia* virus 1 (ABU23712.1); PpV1, *Phaeocystis pouchetii* virus (ABU23715.1); ApMV, *Acathamoeba polyphaga* virus (Q5UQL7.2); EsV1, *Ectocarpus siliculosus* virus (AAK14534.1); FirrV1, *Feldmannia irregularis* virus 1 (AAR26925.1); EhV86, *Emiliania huxleyi* virus 86 (CAI65508.2); EhV99, *E. huxleyi* virus 99 (ABU23713.1); OsV5, *Ostreococcus* virus 5 (ABY27849.1); OtV1, *O. tauri* virus 1 (CAY39653.1); LCDV1, Lymphocystis diseases virus 1(AAC24486.2); FV3, Frog virus 3 (AAT09750.1).

Table 3.3: OLPV and OLV proteins identified in the December 2006 0.1 µm size fraction metaproteome. Peptide sequences by which the proteins were identified are shown in Appendix Table C.1. NSA, normalised spectral abundance; Cov., coverage. *a*Proteins that have some shared peptides; *b*162322406 and 162276024 are protein homologues; *c*A group of proteins containing similar peptides that could not be differentiated by the mass spectral analysis. Only one gene number of that groups is displayed.

| Gene ID | Source | NSA | Accession | Description | Cov. (%) | Peptides (unique) |
|----------------|---------------|------------|------------------|--|-----------------|--------------------------|
| 162322530a | OLPV-1 | 0.000661 | A7U6F0 | Major capsid protein [<i>Phaeocystis pouchetii</i> virus] | 33 | 15 (4) |
| 162322348 | OLPV-1 | 0.000120 | - | - | 11.3 | 2 (2) |
| 162322406b | OLPV-1 | 0.000177 | - | - | 29.4 | 4 (4) |
| 162313481 | OLPV-1 | 0.000010 | YP_002714448 | Leucine rich repeat-containing Miro-like protein [<i>Synechococcus</i> sp. PCC7335] | 3.96 | 2 (2) |
| 162276060 | OLPV-2 | 0.000897 | - | - | 28.9 | 2 (2) |
| 162300260 | OLPV-2 | 0.000226 | - | - | 34.6 | 2 (2) |
| 162276024b | OLPV-2 | 0.000127 | - | - | 16 | 3 (3) |
| 162275992 | OLPV-2 | 0.000098 | NP_048709 | Hypothetical protein PBCV1_A352L [<i>Paramecium bursaria</i> chlorella virus 1] | 16.6 | 2 (2) |
| 162300108 | OLPV-2 | 0.000046 | ZP_01471812 | BNR containing hypothetical protein RS9916_28494 [<i>Synechococcus</i> sp. RS9916] | 7.66 | 5 (5) |
| 162319393a | OLPV-2 | 0.000016 | A7U6F0 | Major capsid protein [<i>Phaeocystis pouchetii</i> virus] | 26.3 | 13 (2) |
| 162300134c | OLPV-1/2 | 0.00010 | AAR21578 | Heat shock protein 70 [<i>Phytophthora nicotianae</i>] | 6.97 | 3 (3) |
| 162286324c | OLPV | 0.000176 | NP_048575 | Hypothetical protein PBCV1_A227L [<i>Paramecium bursaria</i> chlorella virus 1] | 14.7 | 2 (2) |
| OLV9 | OLV | 0.001681 | YP_002122381 | Capsid protein V20 [Sputnik virophage] | 31.1 | 15 (15) |
| OLV8 | OLV | 0.000334 | YP_002122379 | N-term: hypothetical protein V18 [Sputnik virophage] | 19.1 | 8 (8) |
| | | | YP_002122380 | C-term: minor capsid protein V19 [Sputnik virophage] | | |

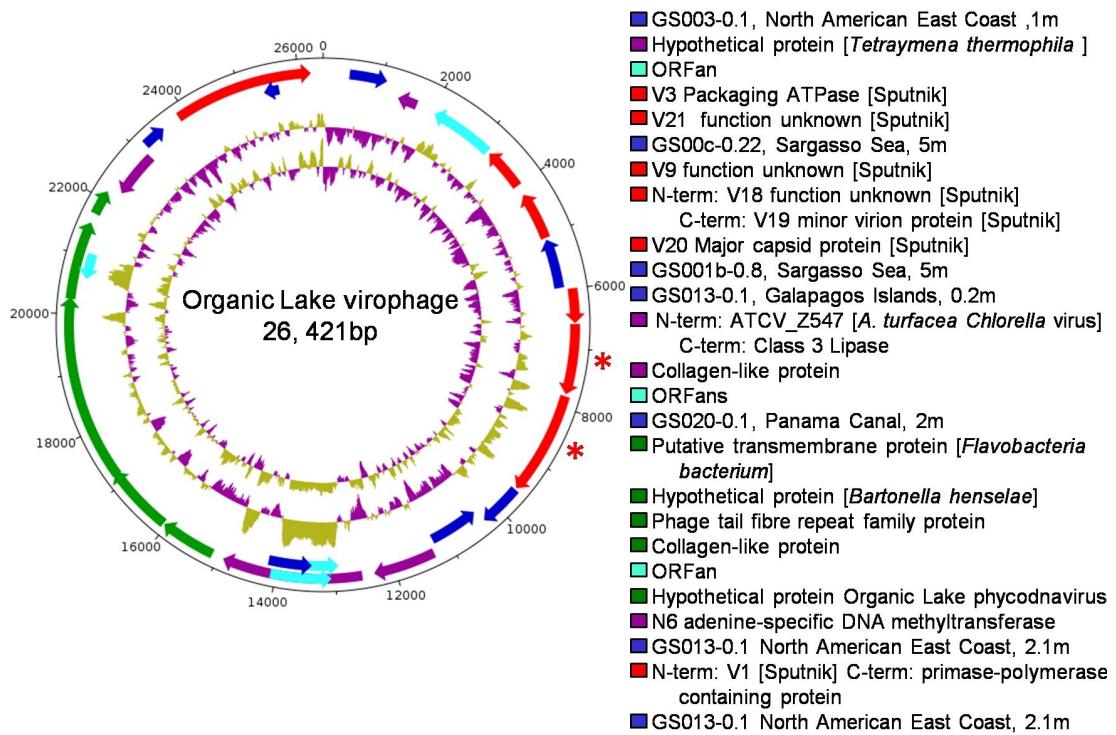


Figure 3.7: Genomic map of Organic Lake virophage. From the outside-in, circles represent, 1) predicted coding sequences on the forward strand, 2) predicted coding sequences on the reverse strand, 3) GC skew, and 4) GC plot. Predicted coding sequences are coloured: Sputnik homologues (red), OLPV homologues (green), non-Sputnik NR homologues (purple), GOS peptide database homologues (blue), and ORFan (cyan). Sequences identified in the metaproteome are marked with an asterisk. Descriptions of the predicted coding sequences from both strands are shown clockwise from position zero.

to the Sputnik genome, and was larger in size (28 kbp vs 18 kbp for Sputnik). Using polymerase chain reaction (PCR) and sequencing, the scaffold was found to represent a complete circular virophage genome. This shows the Organic Lake genome has the same circular topology as the Sputnik genome (La Scola *et al.*, 2008). Virus-like particles resembling Sputnik in size and morphology were identified by TEM (Figure 3.2B).

Table 3.4: Annotation of Organic Lake virophage genome. Top BLASTP matches of predicted coding sequences from the OLV genome compared to OLPV, NR protein database, and CAMERA metagenomic reads ORF peptide database.

| | Gene ID | Start | End | NR (acc, %ID, e-value) | OLPV (geneID, %ID, e-value) | CAMERA (acc, %ID, e-value) |
|---|---------|--------|--------|---|-----------------------------|---|
| † | OLV1 | 460 | 1,077 | - | - | GS003, 0.1, North America East Coast, 1 m (JCVI_PEP_1105157870626/41%/1e-29) |
| | OLV2 | 1,701 | 1,333 | Hypothetical protein [<i>Tetrahymena thermophila</i> SB210] (XP_001029204.1/38% /3e-04) | - | GS012, 0.1, North American East Coast, 13.2 m (JCVI_PEP_1105080106223/42%/1.7e-11) |
| | OLV3 | 3,187 | 2,030 | - | - | - |
| | OLV4 | 3,991 | 3,224 | - | - | - |
| | OLV5 | 5,029 | 4,160 | V3 [Sputnik virophage] (YP_002122364.1/ 39%/4e-24) | - | GS001b, 0.8, Sargasso Sea, 5 m (JCVI_PEP_1105131296011/43%/5e-38) |
| | OLV6 | 5,940 | 5,044 | - | - | GS000c, 0.22, Sargasso Sea, 5m (JCVI_PEP_1105136847382/24%/6e-3) |
| | OLV7 | 5,978 | 6,547 | V9 [Sputnik virophage] (YP_002122370.1/35%/3e-14) | - | GS020, 0.1, Panama Canal, 2 m (JCVI_PEP_1105140820785/26%/1e-12) |
| | OLV8 | 6,574 | 7,740 | N-term: V18 [Sputnik virophage] (YP_002122379.1/27%/9e-05) C-term: V19 [Sputnik virophage] (YP_002122380.1/26%/0.16) | - | GS008, 0.1, North American East Coast, 1 m (JCVI_PEP_1105124194533/32%/6e-8) |
| | OLV9 | 7,791 | 9,518 | V20 [Sputnik virophage] (YP_002122381.1/28%/9e-10) | - | GS033, 0.1, Galapagos Islands, 0.2 m (JCVI_PEP_1105120114513/28%/2e-14) |
| | OLV10 | 9,563 | 10,273 | - | - | GS001b, 0.8, Sargasso Sea, 5 m (JCVI_PEP_1105163928413/61%/5e-4) |
| | OLV11 | 11,210 | 10,317 | - | - | GS013, 0.1, North America East Coast, 2.1 m (JCVI_PEP_1105123792445/39%/9e-37) |
| | OLV12 | 11,284 | 12,324 | N-term: Hypothetical protein ATCV_Z547R [<i>Acanthocystis turfacea</i> chlorella virus 1] | - | GS018, Caribbean Sea, 1.7 m (JCVI_PEP_1105087988121/34%/5.6e-23) |

Continued on next page

Table 3.4 – *Continued from previous page*

| Gene ID | Start | End | NR (acc, %ID, e-value) | OLPV (geneID, %ID, e-value) | CAMERA (acc, %ID, e-value) |
|---------|--------|--------|--|--|--|
| | | | (YP_001427028.1/36%/7e-09) C-term: Lipase class 3 [<i>Bacillus thuringiensis</i> IBL200] (EEM96541.1/27%/1.7e-02) | - | - |
| OLV13 | 12,539 | 14,884 | Collagen-like protein [<i>Bacillus megaterium</i>] (YP_001569009.1/66.67%/4e-03) | - | GS027, 0.1, Galapagos Islands, 2.2m (JCVI_PEP_1105075498120/43%/6.7e-11) |
| OLV14 | 14,023 | 12,905 | - | - | - |
| OLV15 | 13,041 | 14,078 | - | - | - |
| OLV16 | 15,094 | 13,372 | - | - | GS020, 0.1, Panama Canal, 2 m (JCVI_PEP_1105127133835/36%/8.5e-11) |
| 6† | OLV17 | 15,094 | 16,023 Putative transmembrane protein [<i>Flavobacteria</i> bacterium BAL38] (ZP_01734433.1/51%/8e-34) | Lipoprotein Q-like protein (162322444/40%/1e-24) | GS009, 0.1, North American East Coast, 1 m (JCVI_PEP_1105137954859/50%/4e-37) |
| | OLV18 | 16,054 | 17,211 Hypothetical protein BH13620 [<i>Bartonella henselae</i> str. Houston-1] (YP_034083.1/15%/4e-04) | <i>Cyanothece</i> sp. cce_0037-like protein (162322244/65%/2e-33) | GS000c, 0.1, Caribbean Sea, 2 m (JCVI_PEP_1105149563549/39%/2e-26) |
| | OLV19 | 17,168 | 20,278 Phage tail fiber repeat family protein [<i>Trichomonas vaginalis</i> G3] (XP_001296018.1/42%/4e-11) | Lipoprotein Q-like protein (162322444/65%/9e-33) | GS016, 0.1, Caribbean Sea, 2 m (JCVI_PEP_1105149563549/29%/1e-27) |
| | OLV20 | 20,266 | 21,570 Collagen triple helix containing protein A1Q_3499 [<i>Vibrio harveyi</i> HY01] (ZP_01986098.1/69%/6e-04) | Hypothetical protein (162322252/32%/1e-07) | GS033, 0.1, Galapagos Islands, 0.2 m (JCVI_PEP_1105153074955/69%/1e-5) |
| | OLV21 | 21,089 | 20,622 - | - | - |
| | OLV22 | 21,747 | 22,157 - | Hypothetical protein (162322266/56%/5e-31) | GS017, 0.1, Caribbean Sea, 2 m (JCVI_PEP_1105100448171/43%/4e-24) |
| | OLV23 | 23,089 | 22,256 D12 class N6 adenine-specific DNA methyltransferase [" <i>Candidatus Koribacter versatlis</i> " Ellin345] | - | GS002, 0.1, North America East Coast, 1 m (JCVI_PEP_1105085453201/33%/8e-18) |

Continued on next page

Table 3.4 – *Continued from previous page*

| Gene ID | Start | End | NR (acc, %ID, e-value) | OLPV (geneID, %ID, e-value) | CAMERA (acc, %ID, e-value) |
|---------|--------|--------|---|-----------------------------|--|
| OLV24 | 23,174 | 23,560 | - (YP_592471.1/28%/1e-24) | - | GS013, 0.1, North American East Coast 2.1 m (JCVI_PEP_1105132174179/32%/1e-03) |
| OLV25 | 23,889 | 26,219 | N-term: V13 [Sputnik virophage] (YP_002122374.1/34%/5e-31) C-term: Primase-polymerase domain containing hypothetical protein [<i>Ostreococcus lucimarinus</i> CCE9901] (XP_001421479.1/29%/9e-32) | - | GS030, 0.1, Galapagos Islands, 19 m (JCVI_PEP_1105105378071/40%/8e-38) GS013, 0.1, North America East Coast, 2.1 m (JCVI_PEP_1105129419397/51%/8e-71) |
| OLV26 | 25,666 | 25,376 | - | - | GS013, 0.1, North American East Coast, 2.1 m (JCVI_PEP_1105129419399/54%/8e-6) |

Sputnik homologues present in the Organic Lake scaffold included the V20 MCP, V3 DNA packaging ATPase, V13 putative DNA polymerase/primase and others of unknown function (V9, V18, V21 and V32) (Figure 3.7 and (Table 3.3)). The OLV is distinct to Sputnik as proteins share 27–42% amino acid identity (28% MCP identity). OLV proteins include OLV9, the homologue of Sputnik V20 MCP, and OLV8, a fusion of the uncharacterised V18 and minor virion protein V19 from Sputnik (Figure 3.7 and Table 3.4). The large number of homologues, including genes that fulfill essential functions in Sputnik (V20, V3 and V13), indicate that OLV and Sputnik have physiological similarities.

3.4.3 Gene exchange between virophage and phycodnaviruses

As PVs are related to ApMV (Iyer *et al.*, 2006) and are abundant in Organic Lake, it stands to reason that OLPV is the helper of OLV. In the OLV genome, OLV12 is a chlorella virus-derived gene, indicating that gene exchange has occurred between OLV and PVs (the function of OLV12 is discussed below). Similar observations were made for Sputnik, which carries four genes (V6, V7, V12 and V13) in common with the mamavirus, indicative of gene exchange between the viruses and possible co-evolution (La Scola *et al.*, 2008). As the V6, V7, V12 and V13 proteins have been associated with virophage-helper specificity, functional analogues in OLV would have highest identity to proteins from its helper virus, rather than Sputnik.

By comparing OLV and OLPV, a 7,408 bp region was identified in OLV encoding five proteins (OLV17–22) with identity (32–65%) to sequences in both OLPV-1 and OLPV-2 (Figure 3.5, Figure 3.8 and Table 3.4). OLV20 and OLV13 are collagen triple-helix-repeat-containing proteins, analogous to Sputnik collagen-like proteins (V6 and V7) involved in protein–protein interactions in the ApMV virus factory. Sputnik can replicate with either mamavirus or ApMV as a helper, although coinfection rates are higher with the mamavirus. V6 is the only protein with higher identity (69%) to mamavirus than ApMV (42%) (La Scola *et al.*, 2008). Since OLV20 has equivalent identity (63%) with OLPV-1 and OLPV-2, it appears that OLV may be capable of interacting with both OLPV strains. Also within the conserved region, OLV22, is a 141 aa protein of unknown function that only matches sequences from OLPV and the GOS expedition (Table 3.4). Similar to OLV22, Sputnik V12 is a small protein (152 aa) of unknown function with high identity to ApMV, and both may mediate a specific helper–virophage interaction. Other genes in this region of OLV can be mapped to OLPV, including a putative transmembrane protein (OLV17) and paralogous phage tail fibre repeat containing proteins, OLV18 and OLV19. Analogous to the collagen-like proteins, OLV19 and OLV20 probably facilitate interactions between helper and virophage.

OLV12, which is unique to OLV, consists of a C-terminal domain present in conserved hypothetical chlorella virus proteins and an N-terminal domain most closely related to class 3 lipases that may confer OLV selectivity to a PV. OLV12 may function similarly to the Sputnik V15 membrane protein in modifying the ApMV membrane

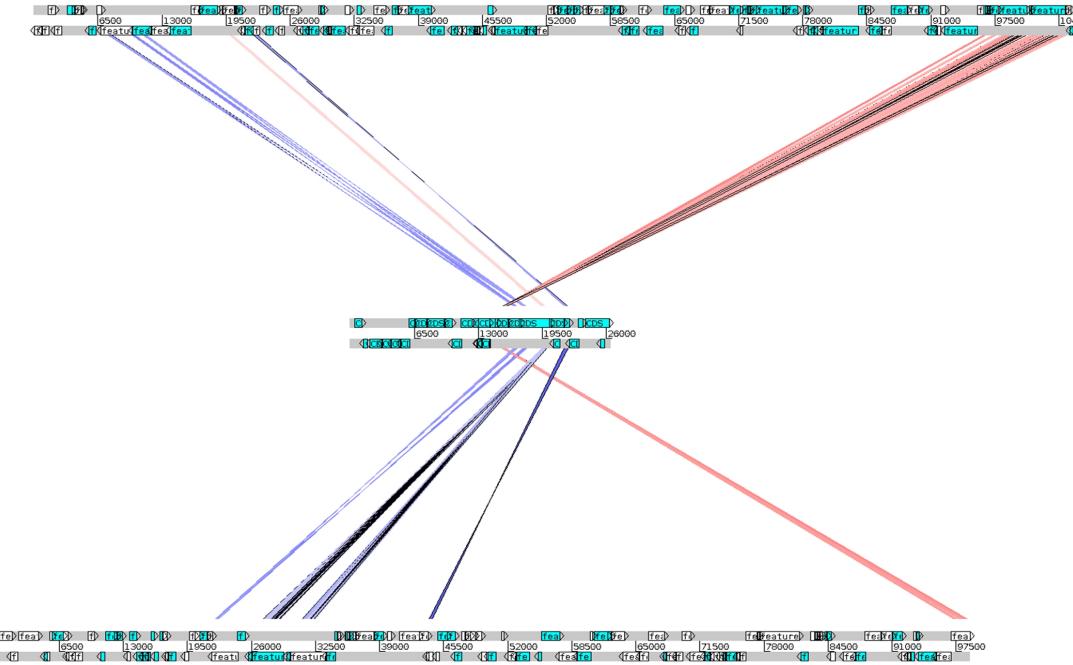


Figure 3.8: Comparison of the location of genes in OLV compared to OLPVs. OLV (centre), OLPV-1 (top), and OLPV-2 (bottom).

(La Scola *et al.*, 2008). The Sputnik V13 consists of a primase domain and SF3 helicase domain related to NCLDV homologues, involved in DNA replication. The helicase domain of OLV25 and V13 are similar, although the primase domain is more similar to a protein from *Ostreococcus lucimarinus*, implying a past association of OLV with a prasinophyte alga host.

Genes unique to OLV point to adaptations specific to its helper-host system. Most notably, OLV possesses a N6 adenine-specific DNA methyltransferase, as does OLPV. In OLPV-1, genes for a bacterial type I restriction modification (RM) system are adjacent to a gene encoding a type I methylase-S target recognition domain protein, and upstream of a DNA helicase distantly related to type III restriction endonuclease (RE) subunits. A large number of chlorella virus genomes have both 5mC and 6mA methylation (Van Etten *et al.*, 1991), and several contain functional RM systems (Nelson *et al.*, 1993). The prototype chlorella virus PbCV-1 possesses REs packaged in the virion for degrading host DNA soon after infection (Agarkova *et al.*, 2006). In contrast to OLV and OLPV, DNA methyltransferases are absent in both Sputnik and APMV, indicating that the N6 adenine-specific DNA methyltransferase has been selected in OLV to reduce endonucleolytic attack mediated by OLPV.

3.4.4 Role of virophage in algal host–phycodnavirus dynamics

The presence of the virophage adds an additional consideration to the microbial loop dynamics. In batch amoeba cultures, co-infection of amoeba with ApMV and Sputnik causes a 70% decrease in infective ApMV particles and a 3-fold decrease in lysis (La Scola *et al.*, 2008). To test how OLV affects OLPV and host population dynamics, OLV

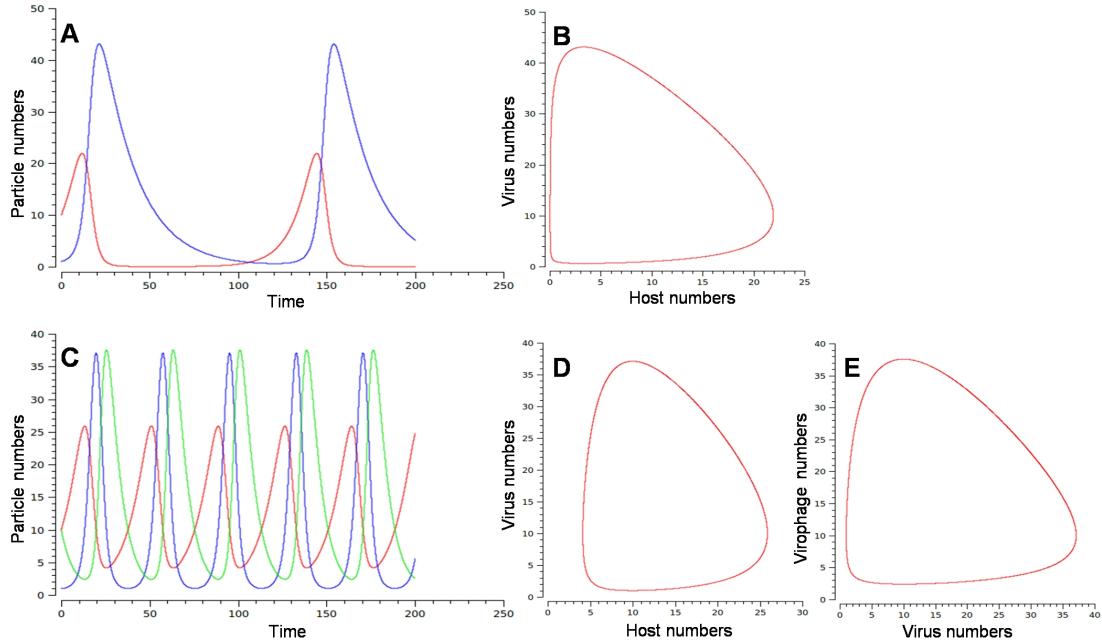


Figure 3.9: Extended Lotka-Volterra models of host-OLPV-OLV population dynamics. **(A)** Time course of host (red line) and OLPV (blue line) populations in the absence of OLV. **(B)** Orbit plot between host and OLPV populations in the absence of OLV with the host and OLPV populations approaching zero during an equilibrium cycle. **(C)** Time course describing the effect of the addition of OLV (green line) on OLPV-host population dynamics as a predator of predator resulting in increased frequency of population oscillations and a higher minimal number of hosts and OLPVs **(D)** compared to in the absence of OLV **(B)**. **(E)** The orbit plot of OLPV and OLV is also shown. Note that the time intervals are arbitrary.

was modelled as an additional predator of a predator in a Lotka-Volterra simulation (Figure 3.9).

The classic Lotka-Volterra model (Lotka, 1910) is based on a pair of first-order, non-linear, differential equations that can be used to describe the periodic oscillation of the populations of a predator and its prey (Volterra, 1926). An example of how predator (virus) populations follows that of its prey (host) populations over time is shown in Figure 3.9A where the populations are at equilibrium. The extended model shown (Figure 3.9C) is based on three equations describing the host (prey), virus (predator) and virophage (predator of predator) interactions. In this model, the effect of virophage is robust, with equilibrium solutions across a wide range of parameter values (Figure 3.9C shows one equilibrium solution). It shows the virus population following that of its host and the virophage population in turn following that of its helper virus or “host”. While the absolute number of hosts do not increase greatly as a result of OLV preying on OLPV, the frequency of host blooms increases in the presence of OLV.

This is due to OLV decreasing the number of infective OLPVs, thereby shortening the recovery time of the host population (Figure 3.9C). This is evident in the orbit plot (Figure 3.9D) as the shift of the orbit away from the axis.

The model reveals that the virophage stimulates the flux of secondary production

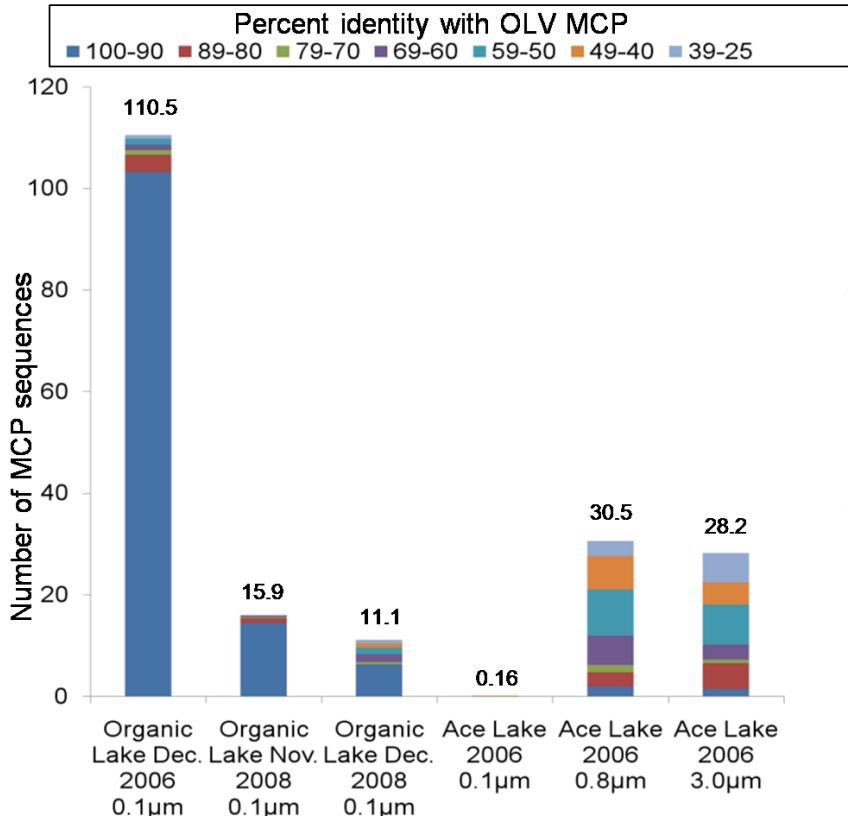


Figure 3.10: Abundance and diversity of virophage capsid proteins in environmental samples. Number of ORFs from metagenomic reads that match to OLV MCP (BLASTP e-value cut-off $1e-5$, abundance normalised to 100,000 reads per sample), and the proportion of virophage capsid types, for the Organic Lake 0.1 µm and Ace Lake 0.1, 0.8 and 3.0 µm fractions.

through the microbial loop by reducing overall mortality of the host algal cell following a bloom, and by increasing the frequency of blooms during the summer light periods. Antarctic lake systems have evolved mechanisms to cope with long light-dark cycles (Lauro *et al.*, 2011) and shortened trophic chains. In Organic Lake and similar systems, a decrease in PV virulence may be instrumental in maintaining stability of the microbial food web. For example, by increasing the frequency of algal blooms more nutrients overall would be released into the available pool for use by heterotrophic bacteria or other primary producers.

3.4.5 Ecological relevance of virophages in aquatic systems

Metagenomic analysis of Organic Lake samples taken two years later in November (when the lake was ice covered) and December 2008 (partially ice-free) revealed sequences with 99% amino acid identity to OLV MCP indicating persistence of OLV in the ecosystem (Figure 3.10 and Table 3.5). In addition, sequences with lower identity (25–90%) were detected, particularly in December, demonstrating Organic Lake virophages are highly diverse but OLV remained the dominant type.

From surface water samples of nearby Ace Lake (meromictic, surface 2% salinity), a large number of sequences were obtained that matched both the OLV MCP (Figure 3.10,

Table 3.5: BLASTP matches for OLV MCP in predicted ORFs of Organic Lake and Ace Lake contigs and CAMERA metagenomic read ORF peptide database. (E-value cut-off 1e-5, alignment length >100aa). Aln., alignment length; Cov., coverage.

| Sample | Size (μm) | Gene ID | Scaffold ID | Id. (%) | Aln. (aa) | E- value | Cov. (×) |
|-----------------|--------------|------------------------|------------------|------------|--------------|-------------|-------------|
| Organic Lake | 0.1 | OLV9 | - | - | - | - | 77.12 |
| December | 0.8 | 176157210 | scf7180000034275 | 98.61 | 575 | 0.0 | 16.03 |
| 2006 | 3.0 | 181703798 | deg7180000108904 | 98.96 | 575 | 0.0 | 48.65 |
| Organic Lake | 0.1 | 192841413 | deg7180000116398 | 99.64 | 555 | 0.0 | 16.03 |
| November | 0.8 | 193037024 | scf7180000086663 | 93.98 | 133 | 2e-61 | 2.5 |
| 2008 | 3.0 | 192638971 | deg7180000028400 | 99.36 | 156 | 1e-76 | 1.86 |
| | | 192955191 | deg7180000024244 | 93.98 | 133 | 1e-61 | 3.10 |
| Organic Lake | 0.1 | 192709908 | scf7180000109753 | 99.01 | 304 | 9e-173 | 4.38 |
| December | | 192709920 | scf7180000109753 | 99.59 | 244 | 1e-120 | 4.38 |
| 2008 | | 192712009 | deg7180000067104 | 54.70 | 117 | 3e-30 | 1.58 |
| | | 192890551 | deg7180000061276 | 36.89 | 122 | 2e-13 | 3.15 |
| | 0.8 | 193060302 | deg7180000053149 | 53.75 | 160 | 6e-43 | 2.30 |
| Ace Lake | 0.1 | 167813925 | scf7180000126822 | 28.86 | 246 | 3e-14 | 2.36 |
| 2006 | | 167858124 | scf7180000129064 | 21.78 | 381 | 5e-10 | 1.94 |
| | | 167891594 | scf7180000136823 | 24.85 | 326 | 2e-04 | 8.15 |
| | | 167875536 | deg7180000086604 | 22.95 | 244 | 6e-04 | 2.21 |
| | 0.8 | 176091445 | deg7180000053588 | 91.61 | 143 | 8e-78 | 3.35 |
| | | 175769103 | deg7180000078701 | 88.24 | 153 | 1e-74 | 1.77 |
| | | 176042318 | deg7180000058177 | 81.77 | 181 | 1e-74 | 2.48 |
| | | 176000635 | deg7180000087166 | 53.39 | 221 | 8e-58 | 2.50 |
| | | 176042707 | deg7180000058207 | 50.78 | 193 | 5e-46 | 2.34 |
| | | 175886340 | deg7180000074162 | 58.90 | 146 | 4e-45 | 2.73 |
| | | 176249679 | deg7180000049481 | 61.94 | 155 | 2e-44 | 2.75 |
| | | 175748439 | deg7180000058552 | 76.79 | 112 | 2e-35 | 2.39 |
| | | 175637390 | deg7180000058712 | 50.91 | 165 | 2e-35 | 2.03 |
| | | 176100822 | deg7180000055966 | 53.38 | 133 | 6e-35 | 1.48 |
| | | 176018109 | deg7180000086684 | 59.68 | 124 | 4e-27 | 1.66 |
| | | 176000624 | deg7180000087165 | 53.85 | 104 | 6e-27 | 1.73 |
| | | 175805608 | deg7180000054222 | 48.60 | 107 | 7e-21 | 1.93 |
| | | 175908895 | deg7180000061971 | 51.91 | 131 | 4e-20 | 3.27 |
| | | 175821062 | deg7180000080443 | 46.46 | 127 | 6e-20 | 1.43 |
| | | 176026419 | deg7180000054364 | 52.59 | 116 | 8e-19 | 1.47 |
| | | 176133336 | scf7180000089989 | 38.36 | 146 | 3e-12 | 1.51 |
| | | 176018257 | deg7180000086719 | 31.21 | 173 | 4e-12 | 1.23 |
| | | 176137412 | deg7180000052688 | 29.37 | 126 | 1e-06 | 2.47 |
| | | 175686880 | scf7180000092161 | 24.00 | 125 | 2e-06 | 2.98 |
| | 3.0 | 175741076 | deg7180000030508 | 85.78 | 232 | 8e-109 | 1.25 |
| | | 175748837 | deg7180000027929 | 55.29 | 170 | 4e-44 | 1.32 |
| | | 175751996 | deg7180000037324 | 51.27 | 158 | 5e-41 | 1.63 |
| | | 175859792 | scf7180000045944 | 30.21 | 288 | 8e-26 | 2.69 |
| Punta Cormorant | 0.1 | JCVI_PEP_1105120114513 | - | 27.84 | 273 | 2e-14 | - |
| hypersaline | | JCVI_PEP_1105100621559 | - | 24.76 | 307 | 9e-10 | - |
| lagoon (GS003) | | JCVI_PEP_1105161421335 | - | 25.61 | 289 | 2e-6 | - |
| Delaware Bay | 0.1 | JCVI_PEP_1105106741177 | - | 24.62 | 264 | 6e-14 | - |
| (GS011) | | JCVI_PEP_1105089715877 | - | 27.16 | 313 | 1e-17 | - |
| Upwelling | 0.1 | JCVI_PEP_1105079267881 | - | 28.23 | 170 | 8e-11 | - |
| (GS031) | | | | | | | |
| Lake Gatun | 0.1 | JCVI_PEP_1105119255775 | - | 26.71 | 149 | 5e-9 | - |
| Panama (GS020) | | | | | | | |

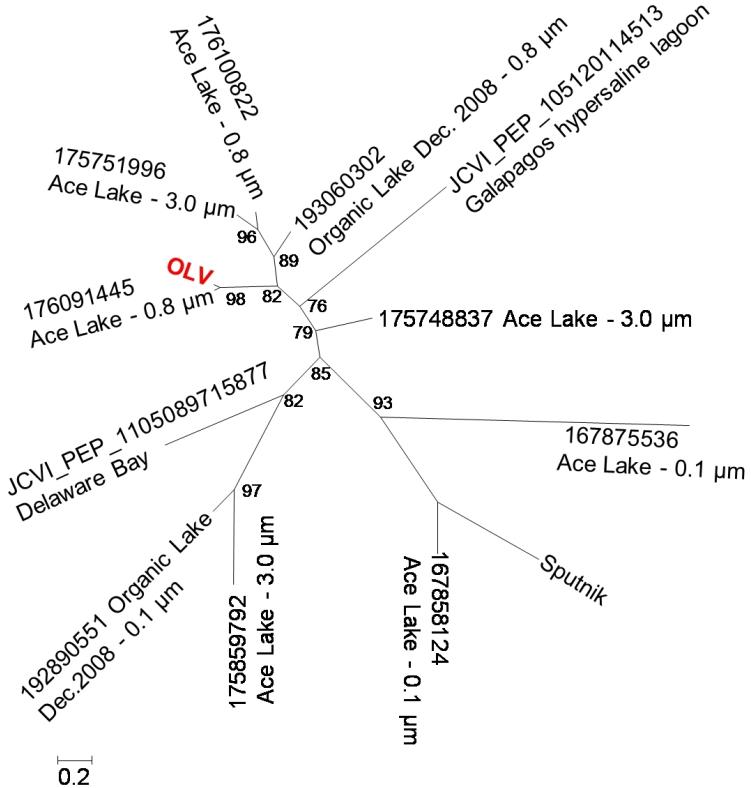


Figure 3.11: Maximum likelihood phylogenetic tree of a conserved 103 aa region of the MCP from Organic Lake, Ace Lake and GOS metagenome data and Sputnik.

Figure 3.11 and Table 3.5) and PVs (Lauro *et al.*, 2011). All Ace Lake size fractions contained matches to OLV MCP, some with high identity (80–100%) and the majority with greater variation (25–80% identity) (Figure 3.11 and Table 3.5). In contrast to Organic Lake where the largest number of matches was to the 0.1 μm size fraction, the majority of Ace Lake sequences were from the larger fractions (Figure 3.10 and Table 3.5). This indicates the Ace Lake virophages (ALVs) were associated with host cells during sampling, or possibly with helper viruses that are larger than the OLPVs.

Extending the OLV MCP search to the GOS data revealed matches (25–28% identity) to sequences from the hypersaline Punta Cormorant Lagoon (Floreana Island, Galapagos), an oceanic upwelling near Fernandina Island (Galapagos), Delaware Bay estuary (NJ, USA), and freshwater Lake Gatun (Panama) (Table 3.5). The phylogenetic analysis of a conserved 103 amino acid region of the MCPs revealed a number of clusters, with Sputnik clustering with virophage sequences from Ace Lake that had low identity (22%) to OLV MCP (Figure 3.11). To improve searches for virophages and better understand their physiology and evolution, it will be valuable to target more genomes (e.g. the ALV 167858124 relative with 40% MCP identity to Sputnik) and determine which genes are core to virophages and what relationship exists between genome complement and MCP identity.

In view of the implications of the virophage modelling (Figure 3.9), the abundance and persistence of OLV in Organic Lake and the presence of diverse virophage signatures

in a variety of lake systems (fresh to hypersaline), an estuary, an ocean upwelling site and a water cooling tower (*Sputnik*), our study indicates that numerous types of virophages exist and play a previously unrecognised role in regulating host–virus interactions and influencing ecosystem function in aquatic environments.

3.5 Acknowledgements

We thank Craig Venter, John Bowman, Louise (Cromer) Newman, Anthony Hull, John Rich and Martin Riddle for providing helpful discussion and logistical support associated with the Antarctic expedition, and Lisa Ziegler for discussion about marine viruses. We acknowledge technical support for computing infrastructure and software development from Intersect, and in particular assistance from Joachim Mai. This work was supported by the Australian Research Council and the Australian Antarctic Division. Funding for sequencing was provided by the Gordon and Betty Moore Foundation to the J. Craig Venter Institute. Mass spectrometric results were obtained at the Bioanalytical Mass Spectrometry Facility within the Analytical Centre of the University of New South Wales. This work was undertaken using infrastructure provided by NSW Government co-investment in the National Collaborative Research Infrastructure Scheme. Subsidized access to this facility is gratefully acknowledged. We thank Jenny Norman from the UNSW Electron Microscopy Unit her assistance in generating images.

Chapter 4

Strategies of carbon conservation and unusual sulphur biogeochemistry in Organic Lake

Co-authorship Statement

A version of this chapter has been submitted as:

Sheree Yau, Federico M. Lauro, Timothy J. Williams, Matthew Z. DeMaere, Mark V. Brown, John Rich, John A.E. Gibson, Ricardo Cavicchioli. Strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline lake. (submitted), 2013.

Contributions to this manuscript by other researchers is as follows.

Research was designed and manuscript edited by Federico Lauro, Tim Williams, John Gibson and Ricardo Cavicchioli. Sample collection and bathymetry was performed by Federico Lauro, Mark Brown, John Rich and Ricardo Cavicchioli. Metagenomic sequence filtering, global assembly and annotation was performed by Matthew DeMaere. Assistance in interpretation of geochemistry provided by John Gibson. Assistance in global taxonomic and functional gene analysis provided by Federico Lauro. Assistance in analysis and interpretation of metabolic pathways provided by Timothy Williams.

Apart from these contributions, I performed all other data analyses, interpretations and drafted the manuscript.

4.1 Abstract

Organic Lake in the Vestfold Hills has the highest reported concentration of dimethylsulphide (DMS) in a natural body of water. To determine the composition and functional potential of the microbial community and learn about the unusual sulphur chemistry in Organic Lake, shotgun metagenomics was performed on size fractionated samples collected along a depth profile. Eucaryal phytoflagellates were the main photosynthetic organisms. Bacteria were dominated by the globally distributed heterotrophic taxa *Marinobacter*, *Roseovarius* and *Psychroflexus*. The dominance of heterotrophic degradation coupled with low fixation potential indicates possible net carbon loss. However, abundant marker genes for aerobic anoxygenic phototrophy, sulphur oxidation, rhodopsins and CO oxidation were also linked to the dominant heterotrophic bacteria and indicate use of photo- and lithoheterotrophy as mechanisms for conserving organic carbon. Similarly, a high genetic potential for the recycling of nitrogen compounds likely functions to retain fixed nitrogen in the lake. dimethylsulphopropionate (DMSP) lyase genes (*dddD*, *dddL* and *dddP*) were abundant indicating DMSP is a significant carbon and energy source. Unlike marine environments, DMSP demethylases (*dmdA*) were less abundant than DMSP lyases indicating that DMSP cleavage is the likely source of the high DMS concentration. DMSP cleavage, photoheterotrophy, lithoheterotrophy and nitrogen remineralisation by dominant Organic Lake bacteria are potentially important adaptations to nutrient constraints. In particular, photo- and lithoheterotrophy reduces the extent of carbon oxidation for energy production allowing more carbon to be used for biosynthetic processes. The study sheds light on how the microbial community in Organic Lake has adapted to the unique physical and chemical properties of this Antarctic lake environment.

4.2 Introduction

Due to the polar light cycle, phototrophic growth in Antarctic environments is relatively high in summer and negligible in winter (Laybourn-Parry *et al.*, 2005) and requires microbial life to survive under long periods under a scarcity of resources. To overcome this limitation, Eucaryotic phytoflagellates in Ace Lake engage in carbon mixotrophy by grazing on bacterioplankton to supplement their carbon requirements in the winter (Laybourn-Parry *et al.*, 2005). Marine heterotrophic bacteria are known to be similarly resourceful by exploiting light energy through photoheterotrophy that includes aerobic anoxygenic photosynthesis (AAnP) or via use of rhodopsins, or lithoheterotrophy such as oxidation of carbon monoxide (Moran and Miller, 2007). Heterotrophic bacteria that can harness energy sources apart from organic carbon can direct a greater proportion of carbon towards growth, which serves to conserve fixed carbon within a closed systems (Moran and Miller, 2007).

The bottom waters of Organic Lake are unusual due to the absence of hydrogen sulphide and the high concentration of the volatile gas dimethylsulphide (DMS) (Deprez *et al.*, 1986; Franzmann *et al.*, 1987b; Gibson *et al.*, 1991; Roberts and Burton, 1993; Roberts *et al.*, 1993). Concentrations of DMS as high as 5,000 nM have been recorded in Organic Lake (Gibson *et al.*, 1991), 100 times the maximum concentration recorded from seawater in the adjacent Prydz Bay and at least 1,000 times that of the open Southern Ocean (Curran *et al.*, 1998). More than forty years ago atmospheric DMS was proposed to have a regulatory effect on global cloud cover as it is a precursor of cloud condensation nuclei (Lovelock and Maggs, 1972; Charlson *et al.*, 1987). However, the first enzymes involved in DMS production were only identified in the last six years (Todd *et al.*, 2007). Rapid progress has been made in this short period and the pathways and organisms involved in DMS transformations have been extensively reviewed (Johnston *et al.*, 2008; Schäfer *et al.*, 2010; Curson *et al.*, 2011b; Reisch *et al.*, 2011; Moran *et al.*, 2012).

The main source of DMS in the marine environment is from the breakdown of dimethylsulphopropionate (DMSP). Eucaryal phytoplankton, in particular diatoms, dinoflagellates and haptophytes, produce large quantities of DMSP, an organo-sulphur compound that is thought to function principally as an osmolyte. DMSP is released due to cell lysis, grazing or leakage and follows two known fates: DMSP cleavage by DMSP lyases (DddD, -L, -P, -Q, -W and -Y) or demethylation by DMSP demethylase (DmdA). Both pathways are associated with diverse microorganisms that can utilize DMSP as a sole carbon and energy source. However, it is only the cleavage pathway that releases volatile DMS that can lead to sulphur loss through ventilation to the atmosphere. The very high levels of DMS in Organic Lake make it an ideal system for identifying the microorganisms and the processes involved in DMS accumulation.

The previous Organic Lake metaproteogenomic study examined viruses from the 0.1 μm fraction of surface water that was collected from Organic Lake in December 2006, and November and December 2008 (Yau *et al.*, 2011) (Chapter 3). In the present study we focused on the cellular population rather than viruses. Our study determined

Table 4.1: Summary of metagenomic data for Organic Lake November 2008 profile.

| ID | Depth (m) | Size (μm) | Trimmed reads | Predicted ORFs (%KEGG matches) | Scaffolds (reads) | >10 kbp scaffolds (reads) | Annotated scaffold ORFs (total ORFs) |
|-------|-----------|------------------------|---------------|--------------------------------|-------------------|---------------------------|--------------------------------------|
| GS374 | 1.7 | 0.1 | 494,573 | 533,468 (31) | 4,318 (63,194) | 5 (771) | 33,262 (83,684) |
| | | 0.8 | 472,635 | 470,949 (52) | 4,161 (126,519) | 68 (17,061) | 37,857 (63,140) |
| | | 3.0 | 158,121 | 158,573 (50) | 2,584 (39,591) | 4 (520) | 18,126 (28,425) |
| GS375 | 4.2 | 0.1 | 541,962 | 556,791 (30) | 4,899 (80,316) | 2 (232) | 35,318 (87,631) |
| | | 0.8 | 472,570 | 492,130 (53) | 5,104 (127,243) | 80 (18,461) | 42,508 (68,366) |
| | | 3.0 | 321,112 | 324,365 (56) | 3,983 (98,102) | 69 (14,713) | 30,938 (51,452) |
| GS376 | 5.7 | 0.1 | 363,280 | 387,528 (25) | 2,342 (39,422) | 6 (1,801) | 21,798 (61,595) |
| | | 0.8 | 484,635 | 448,373 (59) | 6,820 (152,646) | 134 (29,903) | 47,846 (73, 282) |
| | | 3.0 | 290,428 | 292,358 (51) | 3,571 (77,277) | 58 (10,231) | 28,199 (48,910) |
| GS377 | 6.5 | 0.1 | 497,363 | 572,892 (29) | 5,029 (80,520) | 14 (2,711) | 36,685 (92,420) |
| | | 0.8 | 465,381 | 454,018 (51) | 4,202 (129,193) | 57 (17,004) | 43,852 (70,382) |
| | | 3.0 | 187,045 | 211,354 (59) | 2,100 (60,636) | 51 (9,321) | 20,713 (33,497) |
| GS378 | 6.7 | 0.1 | 516,870 | 586,375 (26) | 3,694 (58,618) | 14 (3,422) | 33,243 (96,334) |
| | | 0.8 | 548,253 | 626,115 (57) | 6,957 (161,202) | 136 (32,889) | 56,452 (88,738) |
| | | 3.0 | 202,310 | 219,992 (58) | 2,304 (66,389) | 57 (11,167) | 22,786 (35,034) |

the composition and functional potential of Organic Lake microbiota and, in conjunction with historic and contemporary physico-chemical data, generated an integrative understanding of the whole lake ecosystem.

4.3 Materials and methods

4.3.1 Characteristics of the lake and sample collection

The water level of Organic Lake was measured by surveying as +1.886 m relative to the survey mark (NMV / S / 53) located at 68°27'23.4"S, 78°11'22.6"E. Water was collected from Organic Lake on 10 November 2008 through a 30 cm hole in the 0.8 m thick ice cover above the deepest point in the lake. The sampling hole was established at 68°27'22.2"S, 78°11'23.9"E) following bathymetry measurements constructed on a metric grid. Samples were collected for metagenomics, microscopy and chemical analyses at 1.7, 4.2, 5.7, 6.5 and 6.7 m depths (maximum lake depth 6.8 m).

For metagenomics, lake water was passed through a 20 μm pore size pre-filter, and microbial biomass captured by sequential filtration onto 3.0 μm , 0.8 μm and 0.1 μm pore size 293 mm polyethersulfone membrane filters, and samples immediately preserved in buffer and cryogenically frozen in liquid nitrogen, as described previously (Ng *et al.*, 2010; Lauro *et al.*, 2011). Between 1–2 L of lake water was sufficient to saturate the holding capacity of the filters. DNA was extracted from the filters, samples sequenced using the Roche GS-FLX titanium sequencer, and reads processed to remove low quality bases, assembled and annotated, as previously described (Ng *et al.*, 2010; Lauro *et al.*, 2011). A summary of the 2.4 Gbp of metagenomic data is provided in Table 4.1.

4.3.2 Physical and chemical analyses

An *in situ* profile of pH, conductivity, turbidity, dissolved oxygen (DO) and pressure was measured using a submersible probe (YSI sonde model V6600). A temperature profile was measured using a maximum-minimum mercury thermometer as the YSI probe did not have a capacity to record temperature below -10°C . The 5.7 m sample corresponded to the turbidity maximum and the 6.5 m sample to the turbidity minimum. Conductivity at *in situ* temperature was converted to conductivity at 15°C as described previously (Gibson, 1999). The adjusted conductivity brings the temperature to within a range suitable for estimating practical salinity using the formula of Fofonoff and Millard (1983). Salinity was likely to have been underestimated as it is higher than the range (2–42) for which the conductivity–salinity relation holds. However, the relative difference in salinity between the samples would be accurate.

Density was calculated from the *in situ* conductivity and temperature using the equations described by Gibson *et al.* (1990) and expressed at temperature T as:

$$\sigma_T = (1000 - \text{density}) \text{ kg m}^{-3}$$

Ammonia, nitrate, nitrite, total nitrogen (TN), total dissolved nitrogen (TDN), dissolved reactive phosphorus (DRP), total phosphorus (TP), total dissolved phosphorus (TDP), total organic carbon (TOC), dissolved organic carbon (DOC), total sulphur (TS) and total dissolved sulphur (TDS) were determined by American Public Health Associations Standard Methods at the Analytical Services, Tasmania. Values for dissolved nutrients were measured after filtration through a $0.1 \mu\text{m}$ pore size membrane filter. All other nutrients were measured from water collected after filtration through the on-site $20 \mu\text{m}$ pore size pre-filter.

Ammonia, nitrate, nitrite, DRP, TN, TDN, TP and TDP were measured in a Flow Injection Analyser (Lachat Instruments, Colorado, USA). TOC and DOC were determined in the San++ Segmented Flow Analyser (Skalar, Breda, Netherlands). TS and TDS were analysed in the 730ES Inductively Coupled Plasma–Atomic Emission Spectrometer (Agilent Technologies, California, USA). Principal component analysis PCA was performed using the PRIMER Version 6 statistical package (Clarke and Gorley, 2006) on the normalised physical and chemical parameters.

4.3.3 Epifluorescence microscopy

Water samples collected for microscopy were preserved in formaldehyde (1% v/v). Cells and VLP were vacuum filtered onto 25 mm polycarbonate 0.015 μm pore-size membrane filters (Nuclepore Track-etched, Whatman, GE Healthcare, USA) with a 0.45 μm pore-size backing filter. The 0.015 μm filter was mounted onto a glass slide with ProLong Gold anti fade reagent (Invitrogen, Life Technologies, NY, USA) and 2 μl ($25\times$ dilution in sterile filtered milliQ water $<0.015 \mu\text{m}$) SYBR Gold nucleic acid stain (Invitrogen, Life Technologies, NY, USA). Prepared slides were visualized in an epifluorescence microscope (Olympus BX61, Hamburg, Germany) under excitation with blue light (460–495 nm, emission 510–550 nm). Cell and virus-like particle (VLP) counts were performed on the same filter over 30 random fields of view.

4.3.4 Cellular diversity analyses

Diversity of *Bacteria*, *Archaea* and *Eucarya* was assessed using small subunit ribosomal RNA (SSU) gene sequences. Metagenomic reads that matched the 16S and 18S ribosomal RNA (rRNA) genes were retrieved using METAXA (Bengtsson *et al.*, 2011). Only sequences longer than 200 bp were accepted for downstream analysis.

The Quantitative Insights Into Microbial Ecology (QIIME) pipeline (version 1.4.0) (Caporaso *et al.*, 2010) implementing UCLUST, was used to group SSU sequences into operational taxonomic units (OTUs) at 97% percent identity against the SILVA SSU reference database (release 108) (www.arb-silva.de). SSU sequences that did not cluster with sequences from SILVA were allowed to form new OTUs (no suppression). A representative sequence from each OTU was chosen and classified to the genus level using QIIME implementing the Ribosomal Database Project (RDP) classifier (Wang *et al.*, 2007) trained against SILVA. Assignments were accepted to the lowest taxonomic rank with bootstrap value $\leq 85\%$.

To allow comparison of the relative abundance of taxa, the number of SSU matches per sample filter was normalised to the average number of reads (403,577). Statistical analysis on the relative SSU abundances was performed using the PRIMER version 6 package (Clarke and Gorley, 2006). The SSU counts of each sample filter were aggregated to the genus level and square root transformed to reduce the contribution of highly abundant taxa. A resemblance matrix was computed using Bray-Curtis similarity. The upper mixed zone (1.7, 4.2 and 5.7 m) and deep zone (6.5 and 6.7 m) samples were designated as separate groups and an analysis of similarity analysis of similarity (ANOSIM) performed to test for difference between the two groups. BEST analysis was performed with the abiotic variables: conductivity, temperature, turbidity, DO, pH, TOC, TN, TP, TS, total C:N, total C:P, total N:P, cell counts and VLP counts. The Bio-Env procedure in BEST looks at all the abiotic variables in combination and finds a subset sufficient to best explain the biotic structure. A heat map with bi-clustering dendrogram was generated using R and the package ‘seriation’ (Hahsler *et al.*, 2007) on the normalised square-root transformed SSU counts.

4.3.5 Analysis of functional potential

The relative abundance and taxonomic origin of functional marker genes was used to determine the potential for carbon, nitrogen and sulphur conversions. The ORFs were predicted from trimmed metagenomic reads using METAGENE (Noguchi *et al.*, 2006) accepting those >90 bp in length. Open reading frames ORFs were translated using the standard bacterial/plastid translation table and compared to protein sequences from the KEGG Genes database (release 58) using the basic local alignment search tool (BLAST) (Altschul *et al.*, 1990).

The BLAST output was processed using KEGG Orthology Based Annotation System (KOBAS) version 2.0 (Xie *et al.*, 2011) accepting assignments to KEGG Orthology (KO) groups with e-value $<1e-05$ and rank >5 . KO groups used as functional markers are listed in Supplementary (Table 4.2). Marker enzymes were assigned to taxonomic groups based on the species of origin of the best KEGG Genes match.

Table 4.2: Full list of KEGG orthologs involved in carbon, nitrogen and sulphur conversions that were searched for in the Organic lake metagenome.

| Process | Gene | KO | Notes |
|---------------------------|---|--------|--------------|
| C fixation | ribulose-bisphosphate carboxylase large (<i>cbbL</i>) | K01601 | Calvin cycle |
| | ribulose-bisphosphate carboxylase small (<i>cbbS</i>) | K01602 | Calvin cycle |
| | phosphoribulokinase (<i>prkB</i>) | K00855 | Calvin cycle |
| | ATP-citrate lyase alpha (<i>aclA</i>) | K15230 | rTCA cycle |
| | ATP-citrate lyase beta (<i>aclB</i>) | K15231 | rTCA cycle |
| | citryl-CoA lyase (<i>ccl</i>) | K15234 | rTCA cycle |
| | citryl-CoA synthetase (<i>ccsB</i>) | K15233 | rTCA cycle |
| | carbon monoxide dehydrogenase/acetyl-CoA synthase alpha (<i>cdhA</i>) | K14138 | WL |
| | carbon monoxide dehydrogenase/acetyl-CoA synthase beta (<i>cdhB</i>) | K00190 | WL |
| | | | |
| Respiration | cytochrome C oxidase subunit I (<i>coxI</i>) | K02256 | Eucaryotic |
| | cytochrome C oxidase subunit III (<i>coxIII</i>) | K02262 | Eucaryotic |
| | cytochrome C oxidase subunit I (<i>coxA</i>) | K02274 | Bacterial |
| | cytochrome C oxidase subunit III (<i>coxC</i>) | K02276 | Bacterial |
| Fermentation | L-lactate dehydrogenase (<i>ldh</i>) | K00016 | |
| | pyruvate:ferredoxin oxidoreductase alpha (<i>porA</i>) | K00169 | |
| CO oxidation | pyruvate:ferredoxin oxidoreductase beta (<i>porB</i>) | K00170 | |
| | carbon-monoxide dehydrogenase large (<i>coxL</i>) | K03520 | |
| | carbon-monoxide dehydrogenase medium (<i>coxM</i>) | K03519 | |
| AAnP | carbon-monoxide dehydrogenase small (<i>coxS</i>) | K03518 | |
| | photosynthetic reaction center L (<i>pufL</i>) | K08928 | |
| | photosynthetic reaction center M (<i>pufM</i>) | K08929 | |
| Methanogenesis | coenzyme M methyl reductase (<i>mcrB</i>) | K00401 | |
| | methyl coenzyme M reductase system | K00400 | |
| CH ₄ oxidation | soluble methane monooxygenase | K08684 | |
| | nitrogenase (<i>anfG</i>) | K00531 | |
| N fixation | nitrogenase molybdenum-iron protein alpha (<i>nifD</i>) | K02586 | |
| | nitrogenase iron protein (<i>nifH</i>) | K02588 | |
| | nitrogenase molybdenum-iron protein beta (<i>nifK</i>) | K02591 | |
| | nitric oxide reductase (<i>norB</i>) | K02305 | |
| Denitrification | nitric oxide reductase (<i>norC</i>) | K02305 | |
| | nitrous oxide reductase (<i>nosZ</i>) | K00376 | |
| | periplasmic cytochrome c-552 (<i>nrfA</i>) | K03385 | |
| DNRA | hydroxylamine oxidase (<i>hao</i>) | K10535 | hzo-like |
| | glutamate dehydrogenase (<i>gudB, rocG</i>) | K00260 | |
| N assimilation | glutamate dehydrogenase (NAD(P)+) | K00261 | |
| | glutamate dehydrogenase (<i>gdhA</i>) | K00262 | |
| | assimilatory nitrate reductase | K00360 | |
| | assimilatory nitrate reductase (<i>nasA</i>) | K00372 | |
| | assimilatory nitrate reductase (<i>narG</i>) | K00367 | |
| | glutamine synthetase (<i>glnA</i>) | K01915 | |
| | glutamate synthetase (NADPH/NADH) (<i>gltB</i>) | K00265 | |
| | glutamate synthetase (ferredoxin) (<i>gltS</i>) | K00284 | |
| | ammonia monooxygenase subunit A (<i>amoA</i>) | K10944 | |
| | ammonia monooxygenase subunit B (<i>amoB</i>) | K10945 | |

Continued on next page

Table 4.2 – *Continued from previous page*

| Process | Gene | KO | Notes |
|------------------|--|--------|-------------|
| DSR | ammonia monooxygenase subunit C (<i>amoC</i>) | K10946 | |
| | adenylylsulfatereductase subunit A (<i>aprA</i>) | K00394 | SRB related |
| | adenylylsulfatereductase subunit B (<i>aprB</i>) | K00395 | SRB related |
| | sulfite reductase (<i>dsrA</i>) | K11180 | SRB related |
| | sulfite reductase (<i>dsrB</i>) | K11181 | SRB related |
| ASR | adenylyl sulfate kinase (<i>cysC</i>) | K00860 | |
| | sulfateadenylyltransferase (<i>cysN</i>) | K00956 | |
| | sulfateadenylyltransferase (<i>cysD</i>) | K00957 | |
| S mineralisation | cysteine diogenase (<i>cdoI</i>) | K00456 | |
| | thiosulfate/3-Mercaptopyruvate sulfurtransferase (<i>sseA</i>) | K01011 | |
| DMSP reduction | anaerobic dimethyl sulfoxidereductase A (<i>dmsA</i>) | K07306 | |

Marker genes not represented by a KO group were assessed by BLASTP queries of marker gene sequences with experimentally confirmed function (Table 4.3) against a database of translated ORFs predicted from metagenomic reads. Matches were accepted if the e-value was $<1\text{e}-10$ and sequence identity was within the range shared by homologues of the query sequence(s) (Table 4.3). Matches to marker genes were normalised to 100 Mbp per sample and counted. Normalised frequencies of markers from the same pathway were averaged and those from different pathways were summed.

The same marker genes and BLAST procedure was used to compare the DMSP catabolism and photoheterotrophy potential of Organic Lake with nearby Ace Lake (Lauro *et al.*, 2011), Southern Ocean (SO) (Wilkins *et al.*, 2012) and global ocean sampling (GOS) metagenomes (Rusch *et al.*, 2007). Counts of single copy gene *recA* were also determined to estimate the percentage of genomes containing each marker gene (percentage of marker genes relative to *recA*). Matches to *recA* were accepted with e-value $<1\text{e}-20$ according to the cut-off established by Howard *et al.* (2008). For GOS samples, the BLAST database was generated from peptide sequences retrieved from Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (camera.calit2.net) while the other BLAST databases were produced as for Organic Lake. The total number of trimmed base pairs for GOS samples was estimated by multiplying the number of reads from each sample by the average read length (822 bp) (Rusch *et al.*, 2007).

Marker gene sequences for phylogenetic analysis were clustered using the CD-HIT web server (Huang *et al.*, 2010) at 90% global amino acid identity. A representative sequence from the clusters that resided within a desired conserved region and homologues from cultured strains were used in phylogenetic analyses performed in Molecular Evolutionary Genetic Analysis (MEGA) 5.05 (Tamura *et al.*, 2011) implementing MUSCLE with default parameters (gap opening penalty: -2.9, gap extension penalty: 0). Neighbor-joining was used to compute the phylogenies with a Poisson substitution model, uniform rates of change and complete deletion of alignment gaps. Node support was tested with bootstrap analysis (500 replicates).

Table 4.3: Functional marker gene sequences used in this study as BLAST queries for retrieving homologues in the Organic Lake metagenomes. %ID, minimum amino acid identity for a match to be considered homologous.

| Gene (%ID) | Organism | Accession | Reference |
|-----------------------|--|------------------|--------------------------------------|
| <i>dddD</i> (60) | <i>Marinomonas</i> sp.MWYL1 | ABR72937.1 | Todd <i>et al.</i> (2007) |
| | <i>Pseudomonas</i> sp.J465 | ACY01992.1 | Curson <i>et al.</i> (2010) |
| | <i>Psychrobacter</i> sp.J466 | ACY02894.1 | Curson <i>et al.</i> (2010) |
| | <i>Halomonas</i> sp. HTNK1 | ACV84065.1 | Todd <i>et al.</i> (2010) |
| <i>dddL</i> (45) | <i>Sulfitobacter</i> sp. EE36 | ADK55772.1 | Curson <i>et al.</i> (2008) |
| | <i>Rhodobacter sphaeroides</i> 2.4.1 | ABA77574.1 | Curson <i>et al.</i> (2008) |
| <i>dddP</i> (55) | <i>Roseovarius nubinhibens</i> ISM | EAP77700.1 | Todd <i>et al.</i> (2009) |
| <i>dddQ</i> | <i>Ruegeria pomeroyi</i> DSS-3 | AAV94883.1 | |
| | <i>Roseovarius nubinhibens</i> ISM | EAP76001.1 | |
| | marine metagenome | EAP76002.1 | Todd <i>et al.</i> (2011) |
| | | GOS_7860946 | Todd <i>et al.</i> (2011) |
| | | GOS_2632696 | Todd <i>et al.</i> (2011) |
| | | GOS_2469775 | Todd <i>et al.</i> (2011) |
| <i>dddW</i> | <i>Ruegeria pomeroyi</i> DSS-3 | AAV93771.1 | Todd <i>et al.</i> (2012) |
| <i>dddY</i> | <i>Alcaligenes faecalis</i> | ADT64689.1 | Curson <i>et al.</i> (2011a) |
| <i>dmdA</i> (50) | <i>Ruegeria pomeroyi</i> DSS-3 | AAV95190.1 | Howard <i>et al.</i> (2006) |
| | <i>Pelagibacter ubique</i> HTCC1062 | YP_265671.1 | Howard <i>et al.</i> (2006) |
| <i>rhodopsin</i> | <i>Dokdonia donghaensis</i> MED134 | EAQ40507.1 | Gómez-Consarnau <i>et al.</i> (2007) |
| | <i>Vibrio</i> sp. AND4 | ZP_02194911.1 | Gómez-Consarnau <i>et al.</i> (2010) |
| | <i>Salinibacter ruber</i> DSM 13855 | YP_445623.1 | Balashov <i>et al.</i> (2005) |
| <i>pufL</i> (45) | <i>Roseovarius tolerans</i> | ABK88229.1 | Labrenz <i>et al.</i> (1999) |
| | <i>Congregibacter litoralis</i> KT71 | ZP_01104363.1 | Fuchs <i>et al.</i> (2007) |
| <i>pufM</i> (45) | <i>Roseovarius tolerans</i> | ABK88230.1 | Labrenz <i>et al.</i> (1999) |
| | <i>Congregibacter litoralis</i> KT71 | ZP_01104362.1 | Fuchs <i>et al.</i> (2007) |
| <i>soxB</i> (45) | <i>Sulfurimonas denitrificans</i> DSM 1251 | YP_392780.1 | Sievert <i>et al.</i> (2008) |
| | <i>Thiomicrospira crunogena</i> XCL-2 | ABB42141.1 | Scott <i>et al.</i> (2006) |
| <i>soxA</i> (45) | <i>Sulfurimonas denitrificans</i> DSM 1251 | YP_392780.1 | Sievert <i>et al.</i> (2008) |
| | <i>Thiomicrospira crunogena</i> XCL-2 | YP_390871.1 | Scott <i>et al.</i> (2006) |
| <i>soxC</i> (45) | <i>Sulfurimonas denitrificans</i> DSM 1251 | YP_394569.1 | Sievert <i>et al.</i> (2008) |
| | <i>Thiomicrospira crunogena</i> XCL-2 | YP_390427.1 | Scott <i>et al.</i> (2006) |
| <i>soxD</i> (45) | <i>Sulfurimonas denitrificans</i> DSM 1251 | YP_394568.1 | Sievert <i>et al.</i> (2008) |
| | <i>Thiomicrospira crunogena</i> XCL-2 | YP_390427.1 | Scott <i>et al.</i> (2006) |
| <i>sgr</i> (60) | <i>Sulfurimonas denitrificans</i> DSM 1251 | ABB43898.1 | Sievert <i>et al.</i> (2008) |
| <i>recA</i> | <i>Escherichia coli</i> K12 | P0A7G6.2 | Howard <i>et al.</i> (2008) |

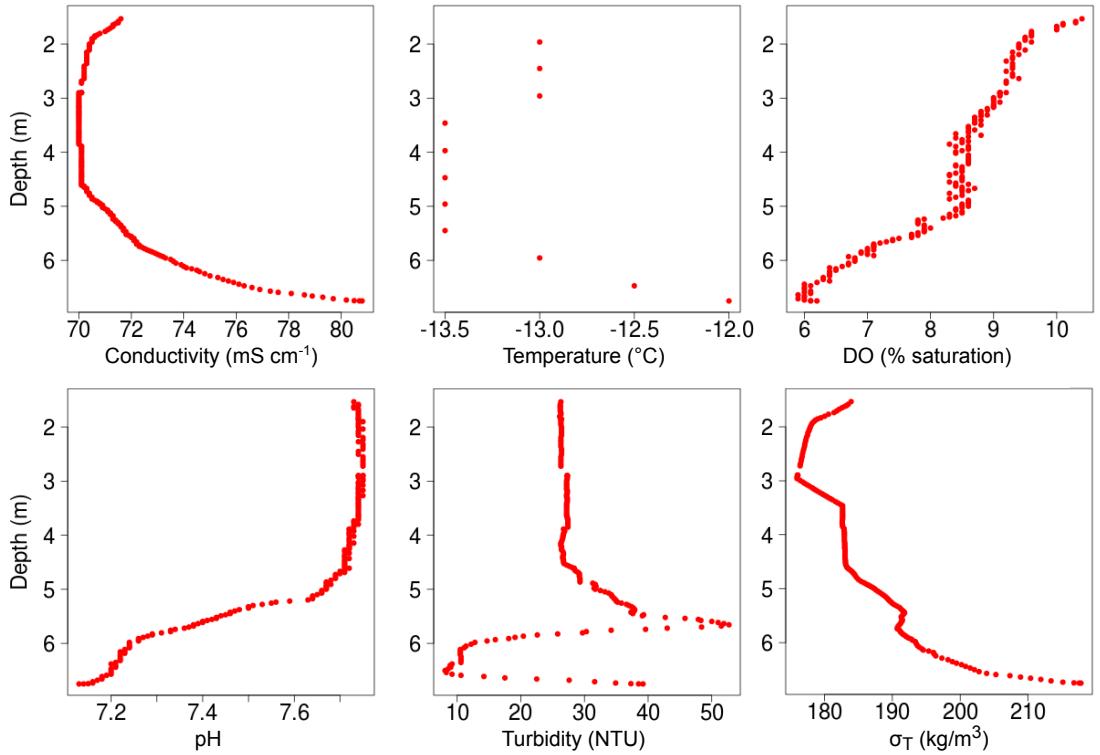


Figure 4.1: Vertical profiles of *in situ* Organic Lake abiotic parameters measured at the deepest point in the lake on 9 November 2008. $\sigma_T = (1000 - \text{density})$ was calculated from temperature and conductivity

4.4 Results and discussion

4.4.1 Abiotic properties and water column structure

In situ physico-chemical profiles (Figure 4.1) measured over the deepest point in the lake (Figure 4.2) determined the existence of two zones: an upper mixed zone above 5.7 m and a suboxic deep zone below 5.7 m (Figure 4.3). The separation of the two zones was indicated by a pycnocline and oxycline starting at 5.7 m. The pH also decreased with DO, likely due to fermentation products such as acetic, formic and lactic acids that have been reported in the bottom waters (Franzmann *et al.*, 1987b; Gibson *et al.*, 1994). The deep zone was not completely anoxic (Figure 4.1). Oxygen may be episodically introduced to bottom waters as a result of currents of cold dense water sinking during surface ice-formation (Ferris *et al.*, 1991). In comparison to meromictic lakes such as Ace Lake that have strong pycnoclines and a steep salt gradient in the anoxic zone, Organic Lake is shallow and has relatively weak stratification (Gibson, 1999).

Samples were collected from the upper mixed (1.7, 4.2 and 5.7 m) and deep (6.5 m and 6.7 m) zones. All nutrients, except for nitrate and nitrite reached maximum concentrations at 6.5 m (Table 4.4) suggestive of a layer of high biological activity above the lake bottom. Consistent with this, cell and VLP counts were highest at 6.5 m. However, turbidity was lowest at this depth demonstrating turbidity was not principally determined by cell density (Figure 4.3). Microscopy images did not show a

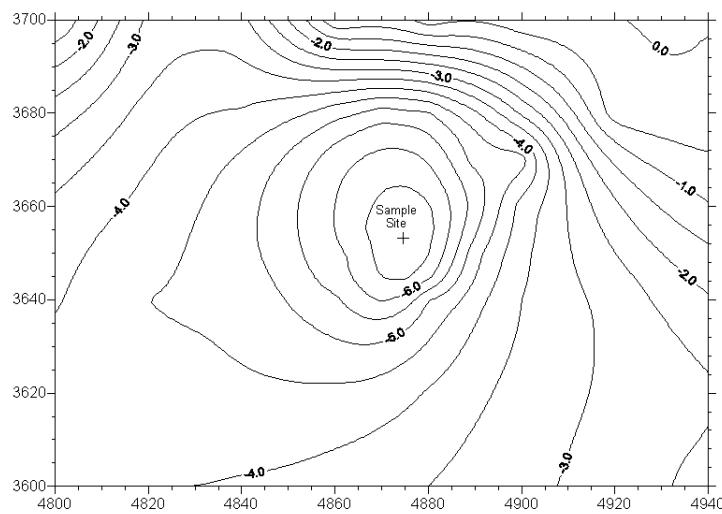


Figure 4.2: Bathymetry of Organic Lake 9 November 2008. Eastings and northings shown are abbreviated metric grid co-ordinates.

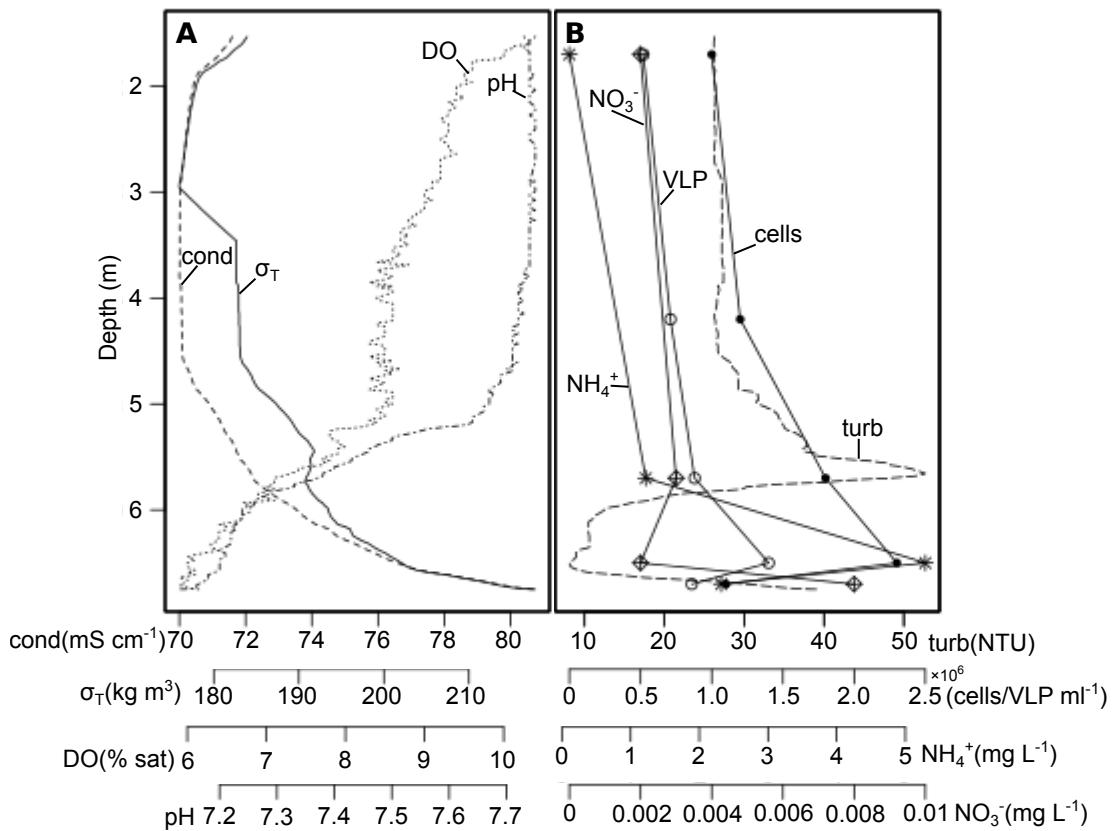


Figure 4.3: Vertical structure of Organic Lake. **(A)** Parameters that varied unimodally with depth showed two zones: an aerobic mixed zone above 5.7 m and a denser suboxic zone below. **(B)** Additional factors that revealed stratification within the deep zone. The peak in concentration at 6.5 m for ammonia was also observed for all other nutrients assayed except nitrate and nitrite, see (Table 4.4) for these values. $\sigma_T = (1000 - \text{density})$; cond, conductivity; DO, dissolved oxygen; turb, turbidity.

Table 4.4: Physico-chemical properties, cell counts and VLP counts of Organic Lake 2008 samples from a vertical profile. ND, data not determined.

| | sample depths (m) | | | | |
|-------------------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| | 1.7 | 4.2 | 5.7 | 6.5 | 6.7 |
| ammonia (mg l ⁻¹) | 0.108 | ND | 1.22 | 5.29 | 2.32 |
| nitrate (mg l ⁻¹) | <0.002 | ND | 0.003 | <0.002 | 0.008 |
| nitrite (mg l ⁻¹) | <0.002 | ND | <0.002 | 0.010 | 0.010 |
| DRP (mg l ⁻¹) | 0.08 | ND | 0.10 | 0.20 | 0.18 |
| TOC (mg l ⁻¹) | 88 | 87 | 110 | 170 | 130 |
| DOC (mg l ⁻¹) | 69 | ND | 97 | 150 | 120 |
| TN (mg l ⁻¹) | 7.70 | 7.50 | 11 | 24 | 13 |
| TDN (mg l ⁻¹) | 0.112 | ND | 1.225 | 5.302 | 2.338 |
| TP (mg l ⁻¹) | 1.5 | 1.4 | 3.0 | 7.6 | 3.7 |
| TDP (mg l ⁻¹) | 0.509 | ND | 0.805 | 4.5 | 2 |
| TS (mg l ⁻¹) | 1010 | 974 | 1020 | 1410 | 950 |
| TDS (mg l ⁻¹) | 996 | ND | 1250 | 1290 | 995 |
| particulate C:N:P (molar ratios) | 49:7:1 | ND | 15:2:1 | 17:3:1 | 15:1:1 |
| dissolved C:N:P (molar ratios) | 350:20:1 | ND | 311:26:1 | 86:10:1 | 155:13:1 |
| practical salinity | 166 | 166 | 172 | 178 | 186 |
| temperature (°C) | -13 | -13.5 | -13 | -12.5 | -12 |
| cells ml ⁻¹ | 1.0±0.4×10 ⁶ | 1.2±0.3×10 ⁶ | 1.8±0.5×10 ⁶ | 2.3±0.8×10 ⁶ | 1.1±0.4×10 ⁶ |
| VLP ml ⁻¹ | 5.2±2.1×10 ⁵ | 7.1±1.3×10 ⁵ | 8.8±3.4×10 ⁵ | 14±3.0×10 ⁵ | 8.6±3.3×10 ⁵ |

shift in cell morphology that could account for the large drop in turbidity (Figure 4.4), which suggests particulate matter primarily contributed to turbidity readings. The low turbidity and peak in cell counts and nutrients at the oxycline at 6.5 m may be caused by an active microbial community degrading particulate matter. This inference is supported by the report of high concentrations of dissolved organic acids and free amino acids in the deep zone (Gibson *et al.*, 1994) as these nutrients are indicative of the breakdown of high molecular weight carbohydrates, lipids and proteins. Furthermore, the C:N and C:P ratios throughout the lake were high compared to the Redfield ratio (Redfield *et al.*, 1963) except at 6.5 m indicating this was the only depth where dissolved nitrogen and phosphorus were not relatively limited (Table 4.4).

Principal component analysis PCA of physico-chemical parameters showed all samples, except the 6.5 m sample, separated with depth along the PC1 axis (Figure 4.5). Accordingly, turbidity, TS and cell density were the strongest explanatory variables for the separation of the 6.5 m sample from the other deep sample, indicating that increased activity at 6.5 m was related to breakdown of particulate matter and sulphur chemistry.

4.4.2 Overall microbial diversity

SSU genes (3,959 reads) that were retrieved from the metagenome data grouped into 983 OTUs. OTUs for *Bacteria* comprised 76.2%, *Eucarya* 16.3% and 7.5% of SSU sequences could not be classified. Only 2 reads, assigned to a deep sea hydrothermal clade of *Halobacteriales* (Supplementary Table S4), were assigned to *Archaea* indicating they were rare in Organic Lake.

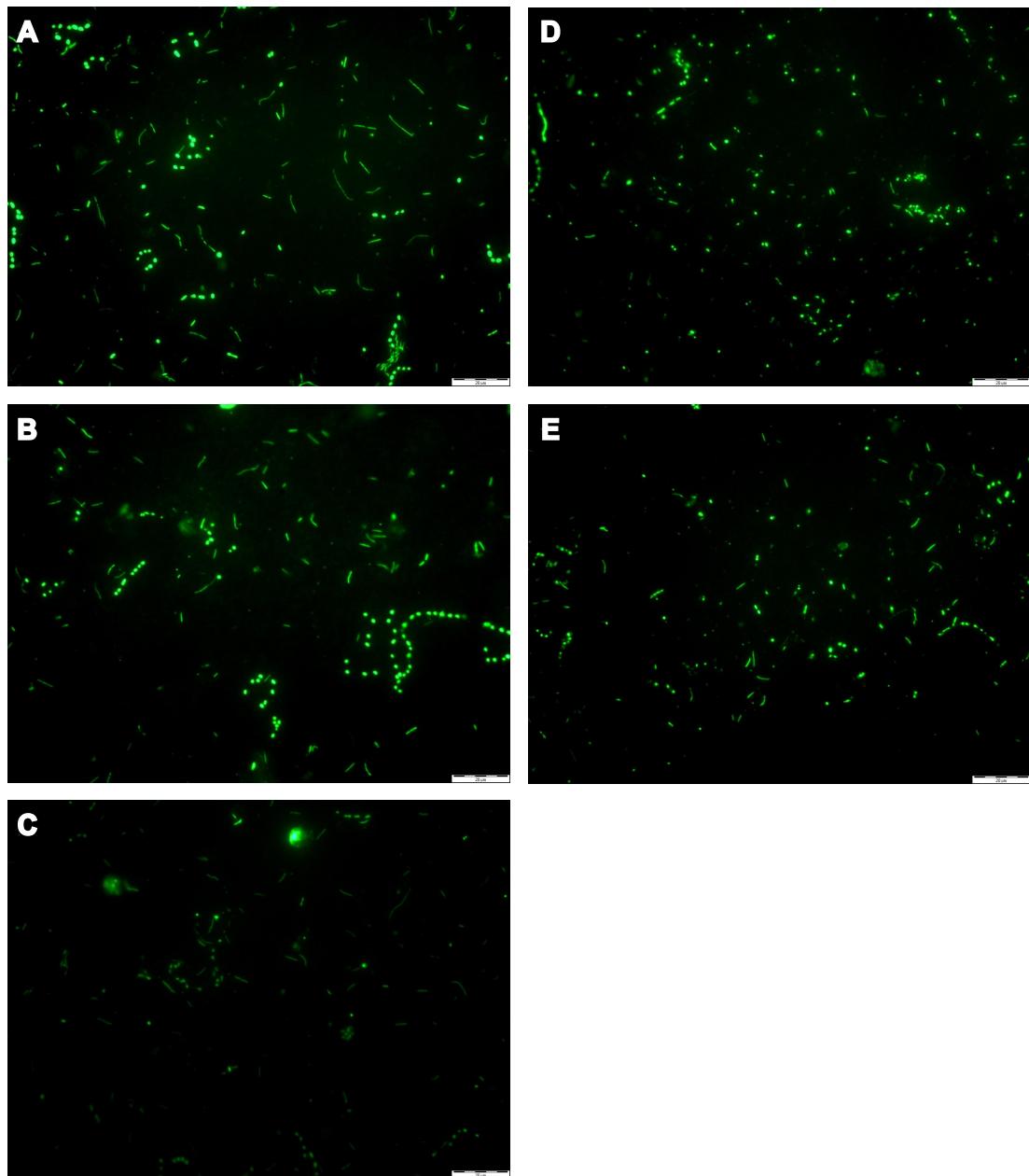


Figure 4.4: Epifluorescence microscopy images of Organic Lake microbiota ($<20\text{ }\mu\text{m}$) onto $0.015\text{ }\mu\text{m}$ polycarbonate membrane and stained with SYBR Gold. (A) 1.7 m, (B) 4.2 m, (C) 5.7 m, (D) 6.5 m, (E) 6.7 m. Scale bar = $20\text{ }\mu\text{m}$.

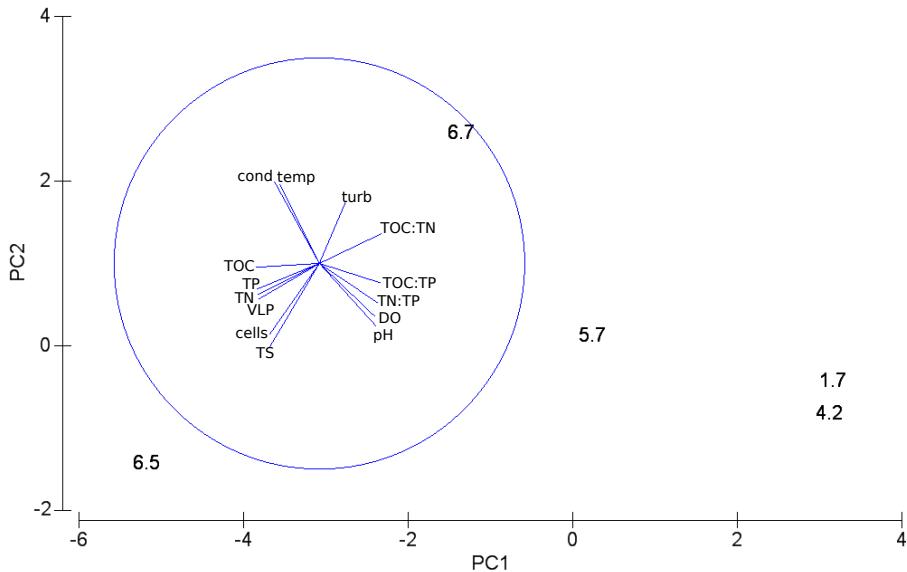


Figure 4.5: PCA of physico-chemical parameters and cell/VLP counts of the Organic Lake profile. Data points are the sampling depths 1.7, 4.2, 5.7, 6.5 and 6.7 m. The overlaid vector diagram shows the relative contributions of the variables to explaining the difference between samples. PC1 explained 74.3% and PC2 explained 14.7% of the variation between samples. cond, conductivity; temp, temperature; turb, turbidity.

The most abundant bacterial classes, *Gammaproteobacteria*, *Alphaproteobacteria* and *Flavobacteria*, were represented by OTUs on all filter sizes at all depths and each consisted of one dominant genus, *Marinobacter*, *Roseovarius* and *Psychroflexus*, respectively (Figure 4.6). Essentially all OTUs for *Cyanobacteria*/chloroplasts were classified as chloroplasts (Figure 4.6), except for three reads that could not be assigned to any lower rank (Supplementary Table S4) indicating free-living *Cyanobacteria* were rare or absent. OTUs for moderately abundant bacterial classes were *Actinobacteria*, *Deltaproteobacteria*, *Epsilonproteobacteria*, and candidate divisions OD1 and RF3. Lower abundance divisions included OTUs for *Bacilli*, *Clostridia*, *Spirochaetes*, *Lentisphaeria*, TM7, *Opitutae*, *Verrucomicrobia*, Bhi80-139, Bd1-5, SR1 and *Chlamydiae* (Figure 4.6).

The dominant eucaryal OTUs were for photosynthetic *Chlorophyta* (green algae) and *Dictyochophyceae* (silicoflagellate algae) (Figure 4.7) principally assigned to the genus *Dunaliella* and the order *Pedinellales*, respectively (Supplementary Table S4). Lower abundance eucaryal OTUs included *Bacillariophyta* (diatoms), *Dinophyceae*, *Fungi* and heterotrophic *Choanoflagellida* and *Ciliophora* (see Supplementary Table S4 for lower taxonomic rank assignments).

4.4.3 Variation of microbial composition according to size and depth

Community composition varied with size fraction and depth. This was supported by seriation analysis that showed samples clustered according to size fraction, and those clusters further separated into upper mixed and deep zone groups (Figure 4.8). A significant difference in genus-level composition between the upper mixed and deep zones was supported by ANOSIM test ($\text{Rho}: 0.53$, significance: 0.1%). Differential

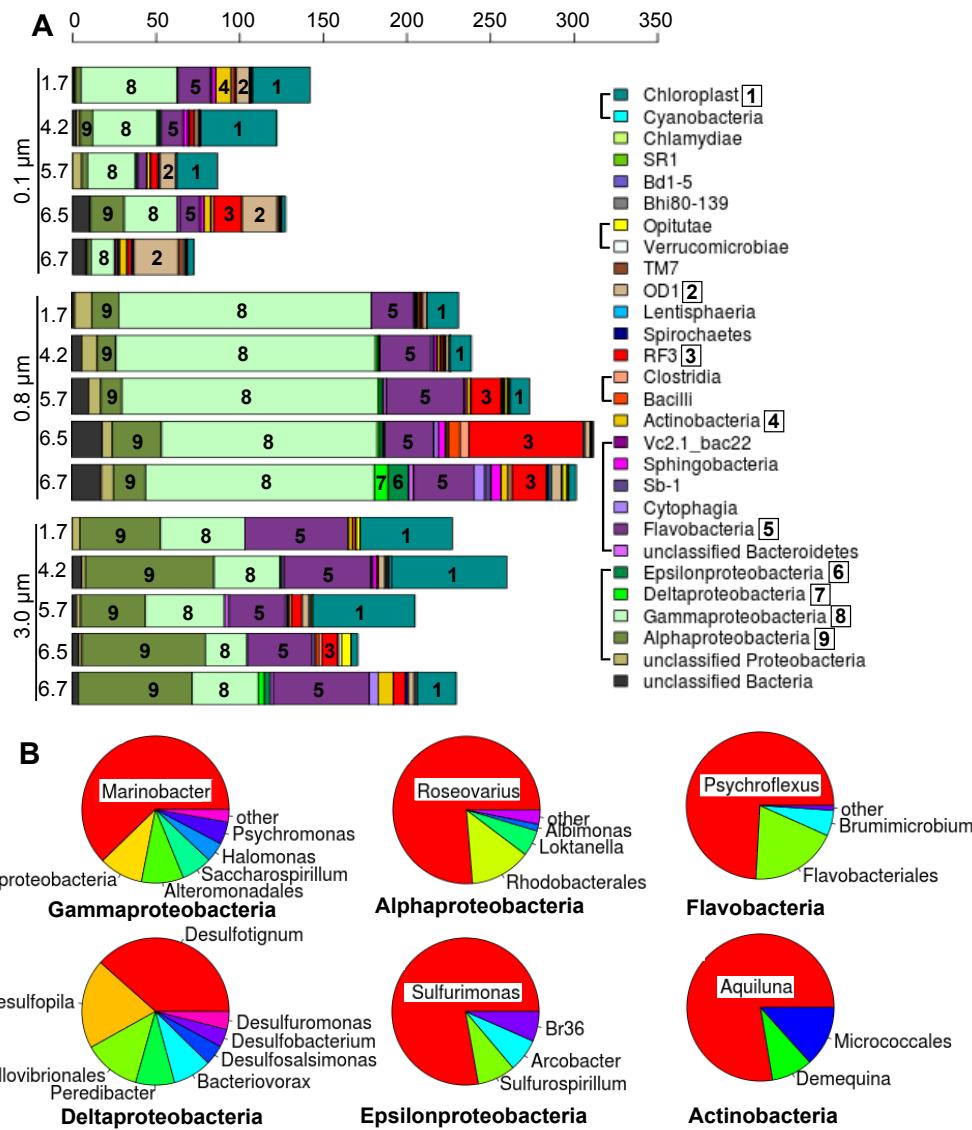


Figure 4.6: Diversity of (A) *Bacteria* from each size fraction (0.1, 0.8 and 3.0 µm) at each sample depth (1.7, 4.2, 5.7, 6.5 and 6.7 m) of Organic Lake aggregated according to class. The x-axis shows counts of SSU normalised to average reads acquired per sample filter. Taxa that belong to the same higher rank are shown grouped with a square bracket in the legend. Abundant taxa are labelled in plot with a number that corresponds to the numbered boxes in the legend. (B) Composition of abundant bacterial classes.

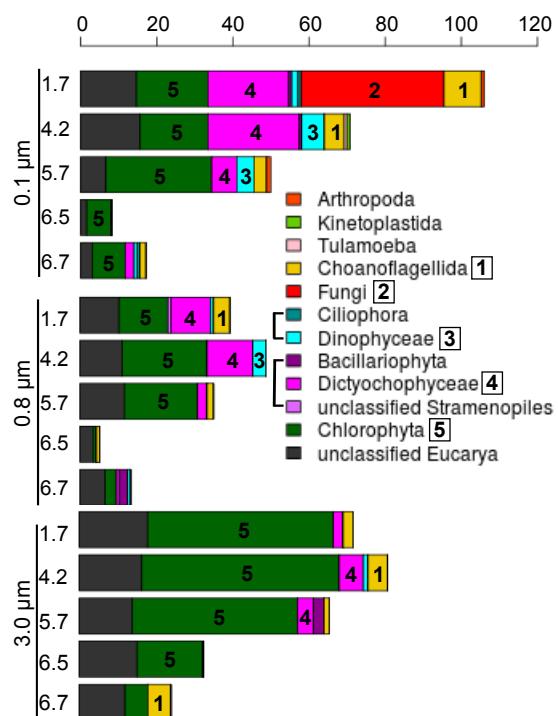


Figure 4.7: Diversity of *Eucarya* from each size fraction (0.1, 0.8 and 3.0 μm) at each sample depth (1.7, 4.2, 5.7, 6.5 and 6.7 m) of Organic Lake aggregated according to class. The x-axis shows counts of SSU normalised to average reads acquired per sample filter. Taxa that belong to the same higher rank are shown grouped with a square bracket in the legend. Abundant taxa are labelled in plot with a number that corresponds to the numbered boxes in the legend.

vertical distribution of taxa is consistent with partitioning of ecological functions in the lake and in association with the physical and chemical data, described functional roles of those taxa.

20–3.0 μm fraction community composition

The upper mixed zone samples had a relatively high OTU abundance of *Dunaliella* chloroplasts and chlorophyte algae consistent with large active photosynthetic organisms concentrating near surface light. They are likely the main source of primary production in Organic Lake and have previously been reported to be the dominant algae (Franzmann *et al.*, 1987b). The SSU sequences for these algae at the bottom of the lake are likely to be due to sedimentation of dead cells or resting cysts.

Psychroflexus OTUs were overrepresented in the surface and 6.7 m samples. Consistent with enrichment on the 3.0 μm filters, *Psychroflexus* (formerly *Flavobacterium*) *gondwanensis* (Bowman *et al.*, 1998) isolated from Organic Lake (Franzmann *et al.*, 1987b) had cells 1.5–11.5 μm in length (Dobson *et al.*, 1991). *Flavobacteria* associate with phytoplankton blooms in the Southern Ocean (Abell and Bowman, 2005a,b; Williams *et al.*, 2012), and have specialized abilities to degrade polymeric substances from algal exudates and detritus (reviewed in Kirchman (2002), (Williams *et al.*, 2012)). It is likely that Organic Lake *Psychroflexus* fills a similar ecological role. In support of this, *Psychroflexus* OTUs cluster with *Dunaliella* chloroplasts in the seriation analysis (Figure 4.8) and *P. gondwanensis* abundance in Organic Lake has been correlated with average hours of sunshine per day indicating population dynamics that is related to summer algal blooms (James *et al.*, 1994). The *Psychroflexus* OTUs in the deep zone are most likely due to sedimentation as *P. gondwanensis* non-motile and strictly aerobic (Dobson *et al.*, 1991).

Roseovarius OTUs were enriched at 4.2 m and 6.5 m suggesting different ecotypes may be present in the upper mixed zone compared to the deep zone. *Roseovarius tolerans*, an isolate from Ekho Lake in the Vestfold Hills, Antarctica has a cell size (1.1–2.2 μm ; (Labrenz *et al.*, 1999)) that would be expected to be captured on the 0.8 μm filter. The *Roseovarius* captured on the 3 μm filter may therefore be a different species, or a strain similar to *R. tolerans* from Ekho Lake that exhibits different growth characteristics (i.e. larger cell size or forms aggregates). A strain of this species from Ekho Lake is capable of microaerophilic growth (Labrenz *et al.*, 1999). Overrepresentation at 6.5 m may therefore be indicative of growth at that depth rather than sedimentation because sinking cells would be more abundant close to the lake bottom at 6.7 m. *Roseovarius* OTUs cluster with *Dunaliella* chloroplast and *Psychroflexus* OTUs in the seriation analysis (Figure 4.8), suggesting that Organic Lake *Roseovarius* may be utilising compounds released from algal-derived particulate matter, or made available by processing of complex organic matter by *Psychroflexus*. *Roseovarius* is a member of the *Roseobacter* clade, which is inferred to have an opportunistic ecology frequently associated with nutrient-replete plankton aggregates, including by-products of flavobacterial exoenzymatic attack (Moran *et al.*, 2007; Teeling *et al.*, 2012). Additionally, the diverse

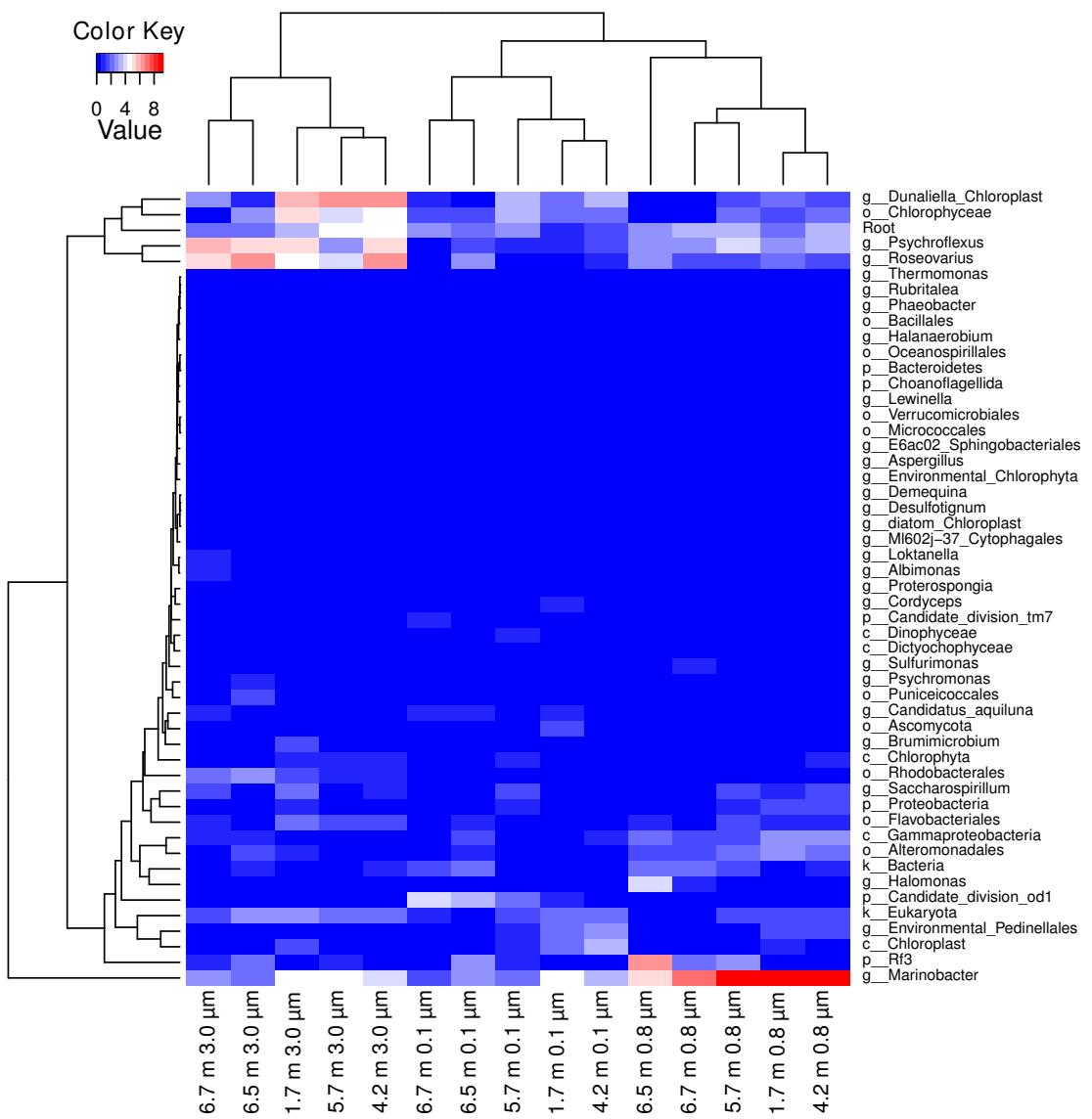


Figure 4.8: Heatmap and biclustering plot of the SSU gene composition in Organic Lake. Samples are shown according to size fraction (0.1, 0.8 and 3.0 μm) and depth (1.7, 4.2, 5.7, 6.5 and 6.7 m). SSU genes were classified to lowest taxonomic rank that gave bootstrap confidence $>85\%$ until the rank of genus. SSU gene counts were normalised and square root transformed. Taxa that comprised $<2\%$ of the sample were not included.

metabolic capabilities of the *Roseobacter* clade include DMSP degradation, AAnP and CO oxidation (reviewed in Wagner-Döbler and Biebl (2006)). All of these capabilities should facilitate growth in both the upper mixed and deep zones of Organic Lake (see 4.4.5).

3–0.8 µm size fraction community composition

On the 0.8 µm filter, OTUs for *Marinobacter* dominated at all depths except 6.5 m. Their capture on this size fraction is consistent with the cell size of isolates (1.2–3 µm) (Gauthier *et al.*, 1992). The genus is metabolically versatile, which likely permits it to occupy the entire water column. *Marinobacter* is heterotrophic and the genus includes hydrocarbon-degrading strains (e.g., Gauthier *et al.* (1992); Huu *et al.* (1996), although deep-sea metal-oxidising autotrophs have also been reported (Edwards *et al.*, 2003). Some isolates are capable of interacting with diatoms (Gärdes *et al.*, 2010) and dinoflagellates (Green *et al.*, 2006). *Marinobacter* isolates from Antarctic lakes are capable of anaerobic respiration using dimethylsulphoxide (DMSO) (Matsuzaki *et al.*, 2006) or nitrate (Ward and Priscu, 1997). Analysis of functional potential linked to *Marinobacter* revealed additional metabolic capabilities potentially related to its dominance in Organic Lake (see Carbon resourcefulness in dominant heterotrophic bacteria and Molecular basis for unusual sulphur chemistry below).

OTUs for RF3 and Halomonas were overrepresented at 6.5 m, and RF3 sequences were more abundant (Figure 4.8). Their relative abundance in the deep zone indicates a role in microaerophilic processes. The majority of RF3 sequences to date are from anaerobic environments including mammalian gut (Tajima *et al.*, 1999; Ley *et al.*, 2006; Samsudin *et al.*, 2011), sediment (Yanagibayashi *et al.*, 1999; Röske *et al.*, 2012), municipal waste leachate (Huang *et al.*, 2005), anaerobic sludge (Chouari *et al.*, 2005; Goberna *et al.*, 2009; Rivière *et al.*, 2009; Tang *et al.*, 2011), a subsurface oil well head (Yamane *et al.*, 2011), and the anaerobic zone of saline lakes (Humayoun *et al.*, 2003; Schmidtova *et al.*, 2009; Bowman *et al.*, 2000b). However, some members have been found in surface waters (Demergasso *et al.*, 2008; Xing *et al.*, 2009; Yilmaz *et al.*, 2012) suggesting not all members are strict anaerobes.

Several *Halomonas* isolates have been sourced from Organic Lake including two described species *Halomonas subglaciescola* and *H. meridiana*, both of which grow as rods with dimensions consistent with capture on this size fraction (Franzmann *et al.*, 1987a; James *et al.*, 1990). Despite these isolates being aerobic, *Halomonas* has been reported to be enriched at the oxycline in Organic Lake (James *et al.*, 1994) indicating *Halomonas* in the lake plays an ecological role in the suboxic zone. This capacity may be linked to the ability of free amino acids and organic acids, which are abundant in the deep zone (Gibson *et al.*, 1994), to stimulate the growth of isolates (Franzmann *et al.*, 1987a).

0.8–0.1 µm size fraction community composition

A large number of eucaryal sequences were evident in the 0.1 µm size fraction. The upper zone was overrepresented by OTUs for *Pedinellales* (silicoflagellate algae) that co-varied with chloroplasts (Figure 4.8). *Pedinellales* have only been detected in Antarctic lakes from molecular studies (Unrein *et al.*, 2005; Lauro *et al.*, 2011) including Organic Lake (Yau *et al.*, 2011) (Chapter 3), and light microscopy studies of Antarctic Peninsula freshwater lakes reported 5–8 µm diameter cells resembling *Pseudopedinella* (Unrein *et al.*, 2005). It is possible that in Organic Lake small (0.80.1 µm) free-living members or chloroplast-containing cyst forms (Thomsen, 2007) exist. However, without evidence to support this (e.g. by microscopy) it seems more likely that the lake sustains a relatively small number of active photosynthetic cells and the sequences detected arise from cysts or degraded cellular material.

OTUs for *Candidatus Aquiluna*, in the Luna-1 cluster of *Actinobacteria* (Hahn *et al.*, 2004; Hahn, 2009) were most abundant at 1.7 m. The genus has small cells (<1.2 µm; (Hahn, 2009), accounting for their concentration on this size fraction. Although originally described in freshwater lakes, the same clade was detected in abundance in Ace Lake (Lauro *et al.*, 2011) and surface Arctic seawater (Kang *et al.*, 2012) demonstrating that they play ecological roles in polar saline systems. In Ace Lake surface waters they were associated with utilisation of labile carbon and nitrogen substrates (Lauro *et al.*, 2011), and in Organic Lake surface waters they probably perform similar functions. The presence of this clade in the deep zone implies a facultative anaerobic lifestyle or sedimented cells.

The bottom of the water column was distinguished by the presence of OTUs for candidate divisions OD1 and TM7. OD1 was more abundant, and its prevalence on this size fraction is consistent with similar findings for size fractionation of ground water (Miyoshi *et al.*, 2005). OD1 is consistently associated with reduced, sulphur-rich, anoxic environments (Harris *et al.*, 2004; Elshahed *et al.*, 2005). OD1 from Zodletone Spring, Oklahoma, was reported to possess enzymes related to those from anaerobic microorganisms (Elshahed *et al.*, 2005). Genomic analyses identified OTUs for OD1 in the anoxic zone of Ace Lake (Lauro *et al.*, 2011). The distribution of OD1 in Organic Lake is consistent with an anaerobic metabolism and potential involvement in sulphur chemistry.

4.4.4 Organic Lake functional potential

To determine the potential for functional processes in Organic Lake, gene markers for carbon, nitrogen and sulphur conversions were retrieved from metagenomic reads. BEST analysis showed that variation in the population structure was significantly correlated (Rho: 0.519, significance: 0.3%) with the abiotic parameters, DO, temperature, TS and TN. The DO gradient has an obvious effect of separating aerobic from anaerobic taxa, and allows oxygen sensitive nitrogen and sulphur processes to occur in the deep zone. Functional potential, taxonomic composition and the physico-chemical data were integrated to infer the carbon, nitrogen and sulphur cycles.

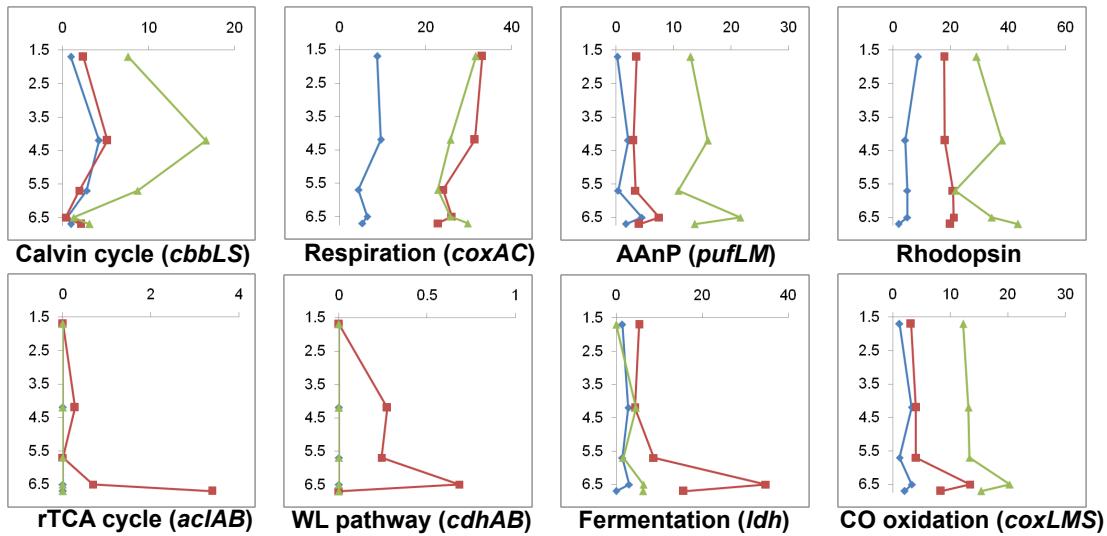


Figure 4.9: Vertical profiles of potential for carbon conversions for each size fraction in Organic Lake. The y-axis shows sample depths (m) and the x-axis shows counts of marker genes normalised to 100 Mbp of DNA sequence. The 0.1, 0.8, 3.0 μm size fractions are shown as blue, red and green, respectively. Counts for marker genes for the same pathway or enzyme complex were averaged and those from different pathways were summed. For marker gene descriptions see Table 4.2 and Table 4.3.

4.4.5 Carbon resourcefulness in dominant heterotrophic bacteria

In both the upper mixed and deep zones, potential for carbon fixation was much lower than for degradative processes, indicating potential for net carbon loss (Figure 4.9). Potential for carbon fixation via the oxygen-tolerant Calvin cycle (Figure 4.9) was originally assessed by presence of the marker genes ribulose-bisphosphate carboxylase oxygenase (RuBisCO) and phosphoribulokinase (*prkB*) (Hügler and Sievert, 2011). The majority of RuBisCO homologues were related to *Viridiplantae* (Table 4.5) supporting the ecological role of green algae as the principle photosynthetic organisms.

RuBisCO was only associated with a small proportion of *Gammaproteobacteria* (Table 4.5), principally from sulphur-oxidising *Thiomicrospira*, indicating some *Gammaproteobacteria* are autotrophs. However, the majority of *prkB* matched to *Gammaproteobacteria* (Table 4.5), predominantly *Marinobacter*. Although deep-sea, iron-oxidising autotrophic members of *Marinobacter* have been isolated (Edwards *et al.*, 2003), all genomes reported for *Marinobacter* have *prkB* but lack RuBisCO genes. Across *Marinobacter* genomes the *prkB* homologue is consistently adjacent to a gene for a putative phosphodiesterase, suggesting that the enzymes expressed by these genes may be involved in a pathway involved in pentose phosphate metabolism unrelated to carbon fixation. Albeit exceptional, this decoupling of *prkB* from RuBisCO involved in carbon fixation (forms I and II), also observed in *Ammonifex* (Hügler and Sievert, 2011), undermines the utility of *prkB* as a marker gene for the Calvin cycle within certain groups. Thus, there is no evidence for autotrophy in Organic Lake mediated by *Marinobacter*.

Evidence for carbon fixation via the reverse tricarboxylic acid (rTCA) cycle was also indicated (Figure 4.9), with genes for ATP citrate lyase (*aclAB*) linked to sulphur-

Table 4.5: Contribution of different taxonomic groups to counts of marker genes involved in carbon conversions. The taxon with the largest contribution to each process is highlighted in blue.

| Taxon | Calvin cycle prkB | Respiration | Fermentation rTCA | WI | CO oxidation AAnP |
|------------------------------|----------------------|-------------|----------------------|------|----------------------|
| <i>Acidobacteria</i> | 0 | 0 | 0.02 | 0 | 0 |
| <i>Actinobacteria</i> | 0 | 0 | 0.64 | 0.23 | 0 |
| <i>Alphaproteobacteria</i> | 0.05 | 0 | 4.84 | 0 | 0 |
| <i>Aquificae</i> | 0 | 0 | 0.06 | 0 | 0 |
| <i>Bacteroidetes</i> | 0 | 0 | 3.42 | 0 | 0 |
| <i>Betaproteobacteria</i> | 0.04 | 0.06 | 0.07 | 0.09 | 0 |
| <i>Chlorobi</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Chloroflexi</i> | 0 | 0 | 0.02 | 0 | 0 |
| <i>Chrysigenetes</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Cyanobacteria</i> | 0.09 | 0 | 0 | 0 | 0 |
| <i>Deferribacteres</i> | 0 | 0 | 0.01 | 0 | 0 |
| <i>Deinococcus-Thermus</i> | 0.01 | 0 | 0.02 | 0 | 0 |
| <i>Deltaproteobacteria</i> | 0 | 0 | 0.09 | 0 | 0 |
| <i>Epsilonproteobacteria</i> | 0 | 0 | 0 | 0.28 | 0 |
| <i>Firmicutes</i> | 0.01 | 0 | 0.01 | 4.90 | 0 |
| <i>Fornicata</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Fusobacteria</i> | 0 | 0 | 0 | 0 | 0.03 |
| <i>Gammaproteobacteria</i> | 0.05 | 12.1 | 9.86 | 1.03 | 0 |
| <i>Nitrospirae</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Planctomycetes</i> | 0 | 0 | 0.02 | 0.08 | 0 |
| <i>Spirochaetes</i> | 0 | 0 | 0 | 0.03 | 0 |
| <i>Thermobaculum</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Thermotogae</i> | 0.01 | 0 | 0 | 0 | 0.17 |
| <i>Verrucomicrobia</i> | 0 | 0 | 0.13 | 0.05 | 0 |
| <i>Crenarchaeota</i> | 0 | 0 | 0 | 0 | 0.01 |
| <i>Euryarchaeota</i> | 0.04 | 0 | 0 | 0 | 0 |
| <i>Alveolata</i> | 0 | 0 | 0.03 | 0 | 0 |
| <i>Euglenozoa</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Opistokonta</i> | 0 | 0 | 0.16 | 0 | 0 |
| <i>Rhodophyta</i> | 0.16 | 0 | 0.03 | 0 | 0 |
| <i>Strameopiles</i> | 0.34 | 0 | 0 | 0 | 0 |
| <i>Viridiplantae</i> | 3.10 | 0.06 | 1.10 | 0 | 0 |

oxidising *Epsilonproteobacteria* (Table 4.5). In general, the rTCA cycle is restricted to anaerobic and microaerophilic bacteria (Hügler and Sievert, 2011), which is consistent with the detection of *Epsilonproteobacteria* in the lake bottom where oxygen is lowest, and the microaerophilic/anaerobic metabolisms characteristic of the group (Campbell *et al.*, 2006). Anaerobic carbon fixation was represented by potential for the Wood-Ljungdahl; or reductive acetyl-CoA (WL) pathway (Figure 4.9). WL-mediated carbon fixation, for which CO dehydrogenase/acetyl-CoA synthase is the key enzyme, was linked to *Firmicutes* and *Deltaproteobacteria* that are known to grow autotrophically using this pathway (Hügler and Sievert, 2011).

Potential for carbon loss by via respiration was indicated by an abundance of cytochrome C oxidase genes (*coxAC*) throughout the water column. In the deep zone, potential for fermentation was greatest at 6.5 m (Figure 4.9) and likely the main biological activity that was occurring at that depth. Fermentation was indicated by the marker gene lactate dehydrogenase (*ldh*). These genes were linked to *Firmicutes* (Table 4.5), which was only present at 6.5 m and represented by the classes *Clostridia* and *Bacilli* (Figure 4.6). As the related candidate division RF3 (Tajima *et al.*, 1999) also has relatively high abundance in this zone (Figure 4.6) (see 0.8–3.0 µm size fraction community composition above), there is circumstantial evidence that RF3 possesses fermentative metabolism and may therefore play an important ecological role in Organic Lake by degrading high molecular weight compounds to organic acids that other organisms can utilize. Assimilation of fermentation products appears to play a greater role in Organic Lake rather than complete anaerobic oxidation involving methanogens or sulphate-reducing bacteria; the former were absent and the latter were present in low abundance (Figure 4.6).

Alphaproteobacteria, predominantly *Roseovarius* (Figure 4.6), were implicated in CO oxidation (Table 4.5), which is used to generate energy for lithoheterotrophic growth (Moran and Miller, 2007), although CO oxidation may also be involved in anaplerotic C fixation (Moran and Miller, 2007). The CO oxidation capacity was at a maximum at 6.5 m (Figure 4.9), and therefore associated with the deep zone *Roseovarius* ecotype of Organic Lake. CO oxidation can function as a strategy to limit oxidation of organic carbon for energy so that a greater proportion can be directed towards biosynthesis (Moran and Miller, 2007).

Photosynthesis reaction center genes *pufLM*, involved in photoheterotrophy via AAnP, were abundant in Organic Lake (Figure 4.9, Table 4.5). These were linked to the *Roseobacter* clade of *Alphaproteobacteria* (Table 4.5), major contributors to AAnP in ocean surface waters (Béjà *et al.*, 2002; Moran and Miller, 2007). This is consistent with the known metabolic potential of bacteriochlorophyll A (BchlA) producing *Roseovarius tolerans* from Ekho Lake (Labrenz *et al.*, 1999). Photoheterotrophy can also be rhodopsin-dependent, with proteorhodopsin (PR) of marine *Flavobacteria* and *Vibrio* previously linked to light-dependent energy generation to supplement heterotrophic growth, particularly during carbon limitation (Gómez-Consarnau *et al.*, 2007, 2010). However, the function(s) of rhodopsins are diverse, and PRs are also hypothesized to

be involved in light or depth sensing (Fuhrman *et al.*, 2008).

Rhodopsin genes were abundant in Organic Lake (Figure 4.9), and were associated with all the dominant Organic Lake aerobic heterotrophic lineages (Figure 4.10). Phylogenetic analysis revealed six well-supported Organic Lake rhodopsin groups (Supplementary Figure S6). All groups had an L or M residue at position 105 (vs the SAR86 PR), denoting tuning to surface green light (Man *et al.*, 2003; Gómez-Consarnau *et al.*, 2007), and is characteristic of oceanic coastal samples (Rusch *et al.*, 2007). Four of the groups clustered with homologues of genera detected in the lake, namely *Marinobacter*, *Psychroflexus*, *Octadecabacter* and “*Ca. Aquiluna*” (Figure 4.10) (Table S4). Another group (SAL-R group) originates from the sphingobacterium *Salinibacter ruber*, which produces xanthorhodopsin (Balashov *et al.*, 2005); it is therefore likely that Organic Lake *Sphingobacteria* (Supplementary Table S4) were the origin of this rhodopsin group.

The most abundant group, OL-R1 (Figure 4.10) had no close homologues from GenBank, but it was abundant on the 3.0 μm fraction and has a distribution suggesting it originates from Organic Lake members of the *Roseobacter* clade (Figure 4.9). All ORFs adjacent to OL-R1 rhodopsin containing scaffolds were related to *Octadecabacter* further supporting their *Roseobacter* clade provenance (Figure 4.11). Genes downstream of OL-R1 were involved in carotenoid synthesis, indicating OL-R1 is a xanthorhodopsin, occurring as a retinal protein or in a carotenoid complex (Balashov *et al.*, 2005).

Photoheterotrophic potential of Organic Lake was compared with other aquatic environments including nearby Ace Lake, SO and GOS expedition samples. The Organic Lake 0.1 μm fraction had the lowest rhodopsin counts and percentage of rhodopsin containing cells of all size-matched samples surveyed (Table 4.6). Non-marine GOS samples from the 0.1 μm fraction have been noted to have lower rhodopsin abundance (Sharma *et al.*, 2008), which was similarly evident from our analysis (Table 4.6). In contrast, the 3.0 μm Organic Lake size fractions had higher rhodopsin counts than Ace Lake and comparable counts to the SO samples, although the percentage of rhodopsin containing cells was still lower than that of the SO. The paucity of rhodopsins in the Organic Lake 0.1 μm fraction is likely due to the lack of SAR11 clade, which is expected to be the main source of rhodopsin genes in Ace Lake and marine samples. This indicates that although Organic Lake has an overall lower frequency of rhodopsin genes compared to sites for which size fraction-matched metagenomes are available, the rhodopsins associated with larger or particle-associated cells are as abundant as in the marine environment.

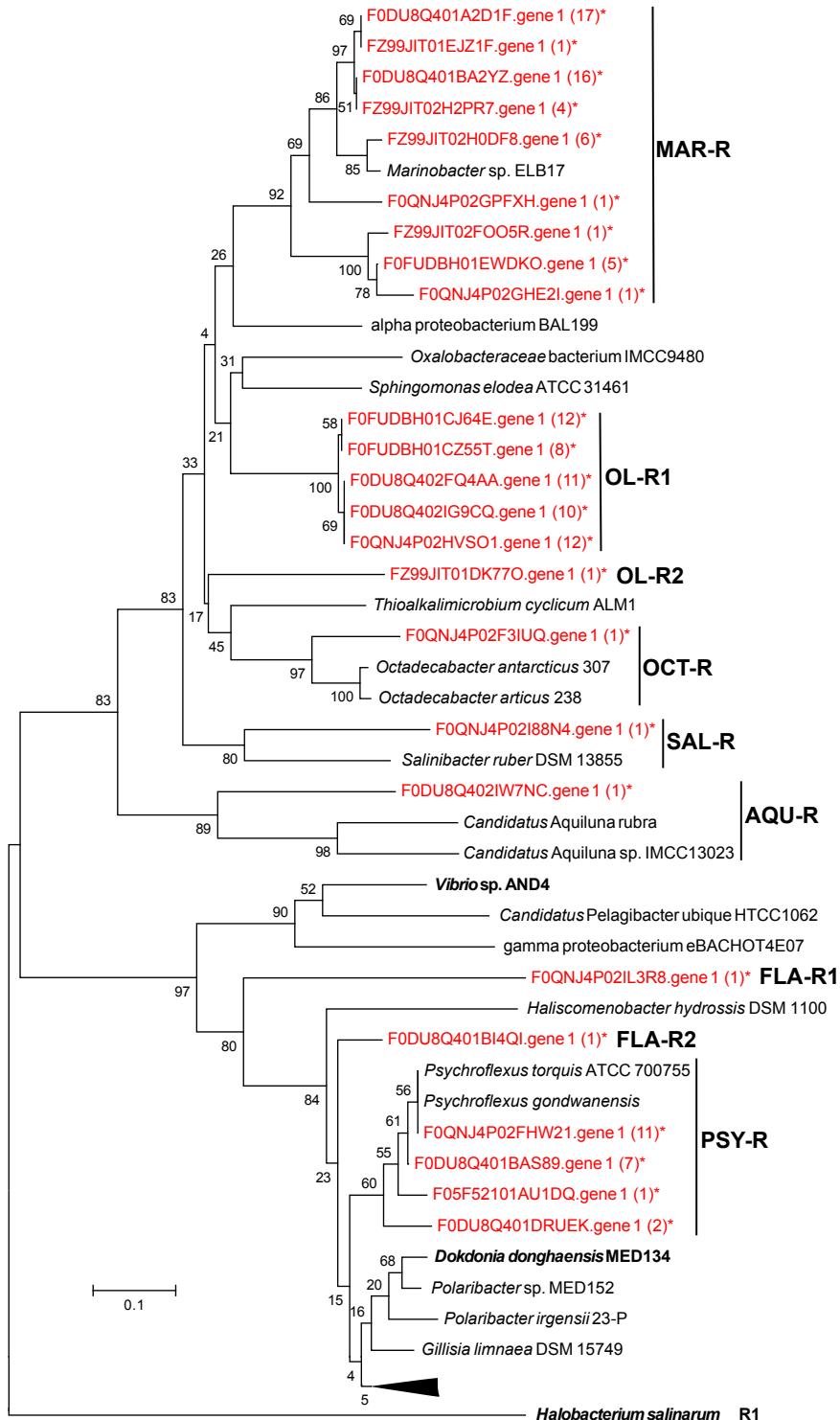


Figure 4.10: Phylogeny of rhodopsin homologs. *Halobacterium salinarum* R1 halorhodopsin was used as an outgroup. The tree was computed from a 78 amino acid region spanning the motif involved in ‘spectral tuning’ using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed activity are shown in bold. Accession numbers from top to bottom are: EAZ99241, EDP63929, EGF32634, ZP_09955974, AEG32267, EDY76405, EDY88259, YP_445623, ACN42850, EIC91904, ZP_02194911, AAZ21446, AAT38609, AEE49633, EAST1907, sequence from John Bowman (personal correspondence), EAQ40507, EAQ40925, EAR12394, EHQ04368 and YP_001689404.

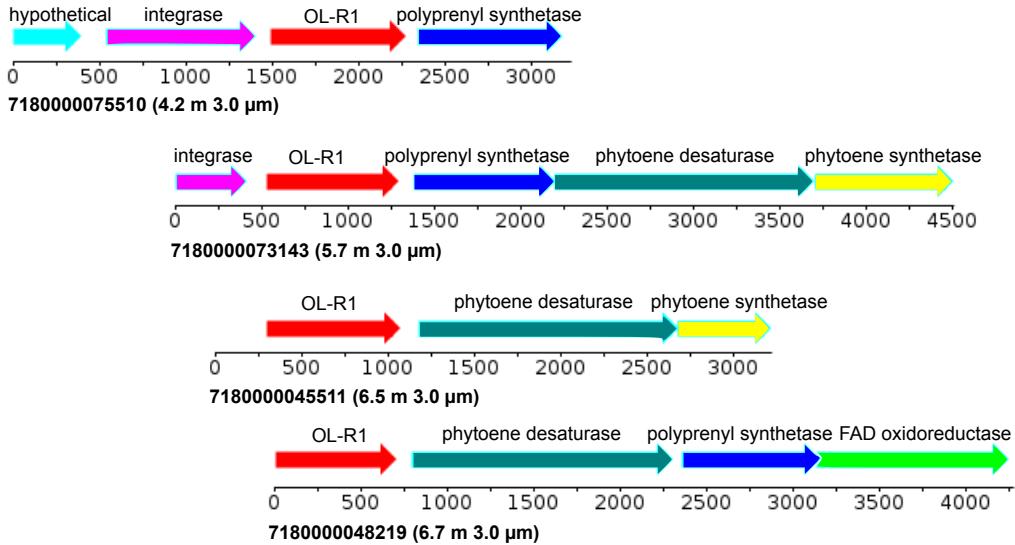


Figure 4.11: Genomic maps of Organic Lake scaffolds containing the OL-R1 rhodopsin homologue. All genes surrounding OL-R1 had best BLAST matches to *Octadecabacter* (*Alphaproteobacteria*) sequences. The scale below shows the number of base pairs. The sample depth and filter from which the scaffold was assembled is shown in parentheses beside the scaffold ID.

Table 4.6: Counts of genes involved in DMSP catabolism and photoheterotrophy in aquatic metagenomes (normalised to 100 Mbp). % = cells containing marker gene. The sample ID for each site is shown in parentheses after the site description. Values marked with an asterisk are >0 but <0.5. The sample with the highest frequency of each marker gene is highlighted in blue. The sample with the highest percentage of cells with each marker is indicated by red font colour. Counts for the following sites are averages of several samples: Ace Lake mixolimnion (GS232, GS231); Southern Ocean SZ (GS349, GS351–GS353, GS356–GS360); Southern Ocean NZ (GS363, GS346, GS364, GS366–GS368); GOS coastal (GS002–GS004, GS007–GS010, GS012–GS016, GS019, GS021, GS027–GS029, GS034–GS036); GOS open ocean (GS017, GS018, GS022, GS023, GS026, GS037, GS047); GOS estuary (GS006, GS011, GS012). Values shown in bold are the highest for that marker gene. SZ, Southern Zone; NZ, Northern Zone; GOS, Global Ocean Sampling.

| Site | Size (μm) | <i>dddD</i> (%) | <i>dddL</i> (%) | <i>dddP</i> (%) | <i>dmdA</i> (%) | Rho. (%) | <i>pufLM</i> (%) | <i>recA</i> |
|-------------------------------|---------------------------|--------------------|--------------------|--------------------|--------------------|-------------|---------------------|-------------|
| Organic Lake 1.7 m (GS374) | 0.1 | 2 (9) | 4 (19) | 0 | 0* (2) | 1 (5) | 0* (1) | 21 |
| | 0.8 | 10 (36) | 10 (39) | 1 (2) | 2 (7) | 5 (20) | 4 (14) | 26 |
| | 3.0 | 11 (50) | 5 (21) | 2 (7) | 9 (43) | 12 (57) | 13 (61) | 21 |
| Organic Lake 4.2 m (GS375) | 0.1 | 5 (34) | 5 (34) | 0 | 1 (10) | 1 (10) | 2 (16) | 14 |
| | 0.8 | 15 (54) | 9 (31) | 0 | 2 (6) | 7 (23) | 3 (11) | 28 |
| | 3.0 | 23 (75) | 2 (8) | 1 (2.5) | 20 (68) | 14 (45) | 16 (53) | 30 |
| Organic Lake 5.7 m (GS376) | 0.1 | 4 (43) | 1 (7) | 0 | 1 (14) | 2 (21) | 0* (4) | 10 |
| | 0.8 | 6 (20) | 9 (32) | 0 | 2 (7) | 6 (22) | 3 (12) | 29 |
| | 3.0 | 19 (68) | 3 (12) | 0 | 13 (47) | 6 (21) | 11 (38) | 28 |
| Organic Lake 6.5 m (GS377) | 0.1 | 10 (51) | 0* (2) | 0 | 3 (15) | 1 (7) | 4 (22) | 20 |
| | 0.8 | 14 (38) | 9 (23) | 1 (2) | 7 (20) | 6 (16) | 7 (20) | 28 |
| | 3.0 | 42 (106) | 5 (13) | 0 | 20 (52) | 6 (16) | 22 (55) | 29 |

Continued on next page

Table 4.6 – *Continued from previous page*

| Site | Size (μm) | <i>dddD</i> (%) | <i>dddL</i> (%) | <i>dddP</i> (%) | <i>dmdA</i> (%) | Rho. (%) | <i>pufLM</i> (%) | <i>recA</i> |
|-------------------------------|---------------------------|--------------------|--------------------|--------------------|--------------------|-------------|---------------------|-------------|
| Organic Lake 6.7 m (GS378) | 0.1 | 1 (7) | 0* (4) | 0 | 0 | 1 (7) | 2 (13) | 13 |
| | 0.8 | 12 (26) | 8 (17) | 0 | 2 (5) | 8 (16) | 4 (9) | 47 |
| | 3.0 | 50 (174) | 5 (17) | 4 (13) | 12 (43) | 12 (43) | 14 (48) | 29 |
| Ace Lake mixolimnion | 0.1 | 0* (2) | 0 | 1 (2) | 15 (56) | 15 (53) | 0* (1) | 28 |
| | 0.8 | 2 (3) | 1 (2) | 0 | 2 (4) | 12 (27) | 3 (12) | 45 |
| | 3.0 | 0 | 0 | 0* (4) | 0 | 5 (42) | 0 | 11 |
| Newcomb Bay (GS235) | 0.1 | 6 (14) | 0 | 3 (7) | 50 (111) | 89 (196) | 0 | 45 |
| | 0.8 | 5 (12) | 0 | 0 | 18 (41) | 55 (123) | 0 | 45 |
| | 3.0 | 0 | 0 | 0 | 2 (17) | 4 (33) | 0 | 11 |
| Southern Ocean SZ | 0.1 | 2 (3) | 0 | 6 (9) | 71 (101) | 98 (139) | 0 | 70 |
| | 0.8 | 3 (6) | 0* (0*) | 5 (12) | 32 (81) | 43 (108) | 0 | 39 |
| | 3.0 | 0* (7) | 0 | 0* (4) | 4 (66) | 5 (84) | 0 | 6 |
| Southern Ocean NZ | 0.1 | 0* (1) | 0 | 5 (7) | 124 (159) | 111 (142) | 1 (1) | 78 |
| | 0.8 | 0* (2) | 0 | 9 (30) | 28 (84) | 35 (107) | 2 (7) | 33 |
| | 3.0 | 0* (3) | 0 | 1 (9) | 7 (54) | 11 (89) | 0* (4) | 12 |
| GOS coastal | 0.1 | 0* (0) | 0 | 5 (6) | 44 (52) | 74 (87) | 5 (6) | 85 |
| GOS open ocean | 0.1 | 0 | 0 | 7 (8) | 45 (50) | 66 (74) | 5 (5) | 90 |
| GOS estuary | 0.1 | 0 | 0 | 1 (1) | 29 (36) | 61 (77) | 2 (3) | 80 |
| GOS embayment (GS005) | 0.1 | 4 (8) | 0 | 6 (12) | 28 (54) | 58 (112) | 3 (6) | 52 |
| GOS Lake Gatun (GS020) | 0.1 | 0 | 0 | 0 | 4 (4) | 48 (53) | 2 (2) | 90 |
| GOS fringing reef (GS025) | 0.1 | 0 | 0 | 0 | 0 | 7 (39) | 0 | 18 |
| GOS warm seep (GS030) | 0.1 | 0 | 0 | 7 (6) | 75 (63) | 83 (69) | 6 (5) | 120 |
| GOS upwelling (GS031) | 0.1 | 0 | 0 | 4 (4) | 81 (77) | 81 (76) | 4 (4) | 106 |
| GOS mangrove (GS032) | 0.1 | 0 | 0 | 2(3) | 24(34) | 25(36) | 1(1) | 71 |
| GOS Punta Cormorant (GS033) | 0.1 | 0 | 11 (15) | 14 (21) | 4 (6) | 31 (43) | 15 (21) | 72 |
| GOS Rangirora Atoll (GS051) | 0.1 | 0 | 0 | 11(15) | 38 (49) | 73 (94) | 3 (4) | 77 |

Counts of *pufLM* genes in the Organic Lake 0.1 μm size fraction were similar to GOS sample, except for Punta Cormorant hypersaline lagoon which had the highest *pufLM* counts and percentage of AAnP cells (Table 4.6). However, the highest overall counts of *pufLM* were from the 3.0 μm size fraction of Organic Lake, likely due to the high proportion of members of the *Roseobacter* clade. Notably, *pufLM* genes were not detected in high abundance in Ace Lake or the SO samples, indicating AAnP is a unique adaptation in Organic Lake among these polar environments. The similarly high abundance of *pufLM* genes in Punta Cormorant hypersaline lagoon indicates AAnP may be advantageous in environments with salinity above marine levels.

The contribution of light-driven energy generation processes to the carbon budget is difficult to infer from genetic potential alone. For example, the relative abundance of

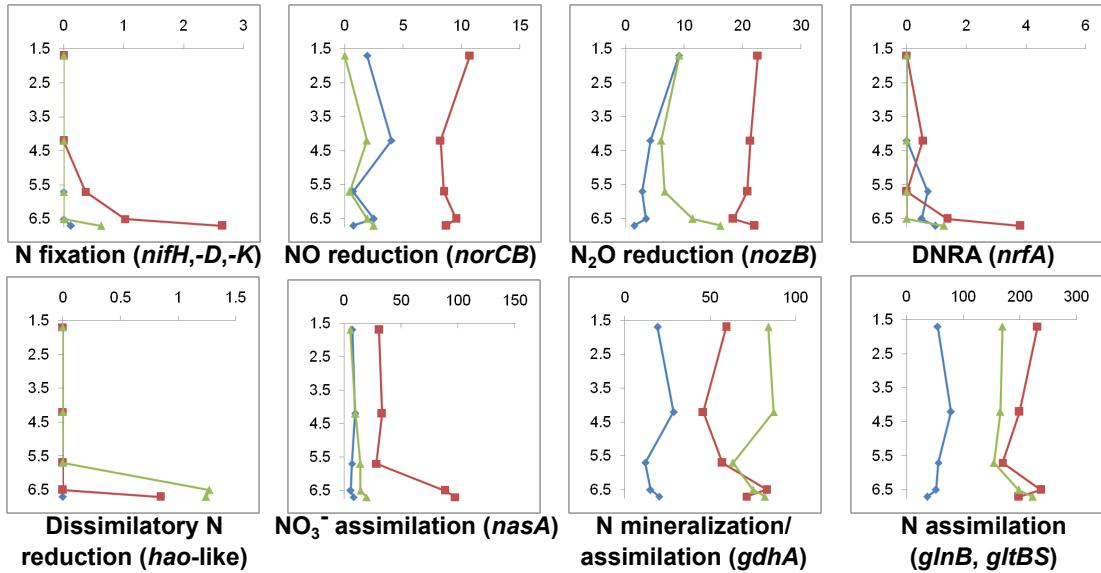


Figure 4.12: Vertical profiles of potential for nitrogen conversions for each size fraction in Organic Lake. The y-axis shows sample depths (m) and the x-axis shows counts of marker genes normalised to 100 Mbp of DNA sequence. The 0.1, 0.8, 3.0 μm size fractions are shown as blue, red and green, respectively. Counts for marker genes for the same pathway or enzyme complex were averaged and those from different pathways were summed. For marker gene descriptions see Table 4.2 and Table 4.3.

AAnP and PR genes in Arctic bacteria has been reported to be the same in winter and summer (Cottrell and Kirchman, 2009). Furthermore, regulation of pigment synthesis is complex; for example, BchlA expression in *R. tolerans* occurs in the dark but is inhibited by continuous dim light (Labrenz *et al.*, 1999). However, it is possible that the apparent negative balance in carbon conversion potential could be ameliorated by photoheterotrophy performed by bacterial groups that are abundant in Organic Lake. In particular, the Organic Lake *Psychroflexus* could play a particular role as it has a PR related to *Dokdonia*, which was shown to function under carbon-limitation (Gómez-Consarnau *et al.*, 2007). Furthermore, detection of higher AAnP potential in Organic Lake than other aquatic environments linked with taxa known to be capable of AAnP, suggests it may have a greater influence in the carbon budget of Organic Lake.

4.4.6 Regenerated nitrogen is predominant in the nitrogen cycle

Nitrogen cycling potential throughout the lake was dominated by assimilation and mineralisation/assimilation pathways (Figure 4.12). Glutamate dehydrogenase (GDH) genes (*gdhA*) were abundant (Figure 4.12), and linked predominantly to *Alpha-* and *Gammaproteobacteria* and to a lesser extent *Bacteroidetes* (Table 4.7). However, the functional significance of the readily reversible GDH depends on its origin; *Bacteroidetes* are likely to use GDH in the oxidative direction for glutamate catabolism (Williams *et al.*, 2012), whereas the use of GDH in the oxidative or reductive directions by *Proteobacteria* is likely to depend upon the source of reduced nitrogen (ammonia vs amino acids). Glutamine synthetase (*glnB*) and glutamate synthase genes (*gltBS*), were pre-

dominantly linked to *Alpha-* and *Gammaproteobacteria* (Table 4.7), indicating the potential for high-affinity ammonia assimilation by these groups in Organic Lake. The high ammonia concentration in the deep zone (Figure 4.3, Table 4.4) would result from a higher rate of mineralisation (ammonification) than assimilation. This is consistent with abundant OTUs for *Psychroflexus* (*Bacteroidetes*) in this zone, and due to either turnover of organic matter or lysis of *Bacteroidetes* cells after sedimentation in anoxic water. In addition, the gene for ammonia-generating nitrite reductase (*nrfA*) was linked to *Bacteroidetes* and *Planctomycetes* (Table 4.7), indicating ammonia may also be produced by these putative aerobic heterotrophs. Overall, the data suggest that ammonia is actively assimilated in the aerobic upper mixed zone, but is permitted to accumulate in the anaerobic deep zone.

Potential for nitrogen conversions typically found in other aquatic environments was greatly reduced in Organic Lake. There was a very low potential for nitrogen fixation that was confined to the deep zone (Figure 4.12) and principally linked to anaerobic *Epsilonproteobacteria* (Table 4.7). This diazotrophic potential may not be realized by *Epsilonproteobacteria*, given the high ammonia concentration present in the deep zone. No ammonia monooxygenase genes (*amoA*) were detected. The potential for ammonia oxidation was only represented by hydroxylamine/hydrazine oxidase-like (*hao*) genes, which were in low abundance and linked to *Deltaproteobacteria* (Table 4.7). *hao* genes are present in non-ammonia-oxidising bacteria (Bergmann *et al.*, 2005), and those from Organic Lake belong to a family of multiheme cytochrome c genes present in sulphate-reducing *Deltaproteobacteria* that have no proven role in ammonia oxidation. In the genomes of sulphate-reducing *Deltaproteobacteria* the *hao* gene is invariably situated adjacent to a gene for a NapC/NirT protein, which suggests a role in dissimilatory nitrate reduction. Collectively these data indicate an inability for nitrification to occur in the upper mixed zone and no potential for ammonia loss in the deep zone.

Denitrification genes (*norCB* and *nozB*) and genes for nitrate assimilation (*nasA*) were present throughout the water column (Figure 4.12) and were linked primarily to *Gammaproteobacteria* (Table 4.7). Low nitrate and nitrite in the deep zone (Figure 4.3, Table 4.4) indicates oxidized nitrogen has been depleted by dissimilatory or assimilatory reduction by heterotrophic *Gammaproteobacteria*. Denitrification genes are phylogenetically widespread and usually induced by low oxygen or oxidized nitrogen species (Kraft *et al.*, 2011) and thus expected to be active in the deep zone or oxycline. However, denitrification may be inhibited even if conditions appear appropriate. For example, in Lake Bonney, Antarctica, denitrification occurs in the west lobe, but not in the east lobe of the lake despite the presence of anoxia, nitrate and denitrifying *Marinobacter* species (Ward and Priscu, 1997; Ward *et al.*, 2005). Moreover, in the absence of nitrification, denitrification and nitrate assimilation would be limited by the lack of potential to re-form oxidized nitrogen. The preponderance of assimilation/mineralisation pathways geared towards reduced nitrogen appears to reflect a “short circuit” of the typical nitrogen cycle that would conserve nitrogen in a largely closed system. Hence, the predominant nitrogen source is regenerated fixed nitrogen. Similar findings were

Table 4.7: Contribution of different taxonomic groups to counts of marker genes involved in nitrogen conversions. The taxon with the largest contribution to each process is highlighted in blue.

| Taxon | <i>N</i> fixation | NO reduction | <i>N</i> ₂ O reduction | DNRA | <i>hao</i> | <i>N</i> mineralisation | <i>NO</i> ₃ ⁻ assimilation | <i>N</i> assimilation |
|------------------------------|-------------------|--------------|-----------------------------------|------|------------|-------------------------|--|-----------------------|
| <i>Acidobacteria</i> | 0 | 0 | 0 | 0 | 0 | 0.07 | 0 | 0.08 |
| <i>Actinobacteria</i> | 0 | 0 | 0 | 0.03 | 0 | 0.32 | 0 | 5.41 |
| <i>Alphaproteobacteria</i> | 0.01 | 0.12 | 0 | 0 | 0 | 6.39 | 5.49 | 49.4 |
| <i>Aquificae</i> | 0 | 0 | 0.26 | 0 | 0 | 0 | 0 | 0.06 |
| <i>Bacteroidetes</i> | 0 | 0 | 3.00 | 0.27 | 0 | 3.90 | 0.03 | 15.5 |
| <i>Betaproteobacteria</i> | 0 | 0.03 | 0 | 0 | 0 | 0.06 | 0.41 | 19.2 |
| <i>Chlorobi</i> | 0.03 | 0 | 0 | 0 | 0 | 0.2 | 0 | 0.31 |
| <i>Chloroflexi</i> | 0 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0.03 |
| <i>Chrysigenetes</i> | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0.06 |
| <i>Cyanobacteria</i> | 0 | 0 | 0 | 0 | 0 | 0.29 | 0 | 0.10 |
| <i>Deferrribacteres</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>Deinococcus-Thermus</i> | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0.11 |
| <i>Deltaproteobacteria</i> | 0.04 | 0.01 | 0 | 0.07 | 0.22 | 0.23 | 0 | 0.58 |
| <i>Epsilonproteobacteria</i> | 0.32 | 0 | 0 | 0 | 0 | 0.05 | 0 | 1.49 |
| <i>Firmicutes</i> | 0.03 | 0 | 0 | 0.03 | 0 | 0.70 | 0 | 3.16 |
| <i>Fornicata</i> | 0 | 0 | 0 | 0 | 0 | 0.02 | 0 | 0 |
| <i>Fusobacteria</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 |
| <i>Gammaproteobacteria</i> | 0 | 3.91 | 8.28 | 0 | 0 | 4.75 | 14.1 | 50.6 |
| <i>Nitrospirae</i> | 0 | 0 | 0 | 0.03 | 0 | 0.01 | 0 | 0 |
| <i>Planctomycetes</i> | 0 | 0.01 | 0 | 0.16 | 0 | 0 | 0 | 0.26 |
| <i>Spirochaetes</i> | 0 | 0.01 | 0 | 0 | 0 | 0.11 | 0 | 0.15 |
| <i>Thermobaculum</i> | 0 | 0 | 0 | 0 | 0 | 0.10 | 0 | 0 |
| <i>Thermotogae</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>Verrucomicrobia</i> | 0 | 0 | 0.17 | 0.03 | 0 | 0.25 | 0 | 0.82 |
| <i>Crenarchaeota</i> | 0 | 0 | 0 | 0 | 0 | 0.02 | 0 | 0 |
| <i>Euryarchaeota</i> | 0 | 0 | 0 | 0 | 0 | 0.12 | 0.09 | 0.10 |
| <i>Alveolata</i> | 0 | 0 | 0 | 0 | 0 | 0.18 | 0 | 0.03 |
| <i>Euglenozoa</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>Opistokonta</i> | 0 | 0 | 0 | 0 | 0 | 0.15 | 0 | 0.13 |
| <i>Rhodophyta</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| <i>Strameopiles</i> | 0 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0.15 |
| <i>Viridiplantae</i> | 0 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0.35 |

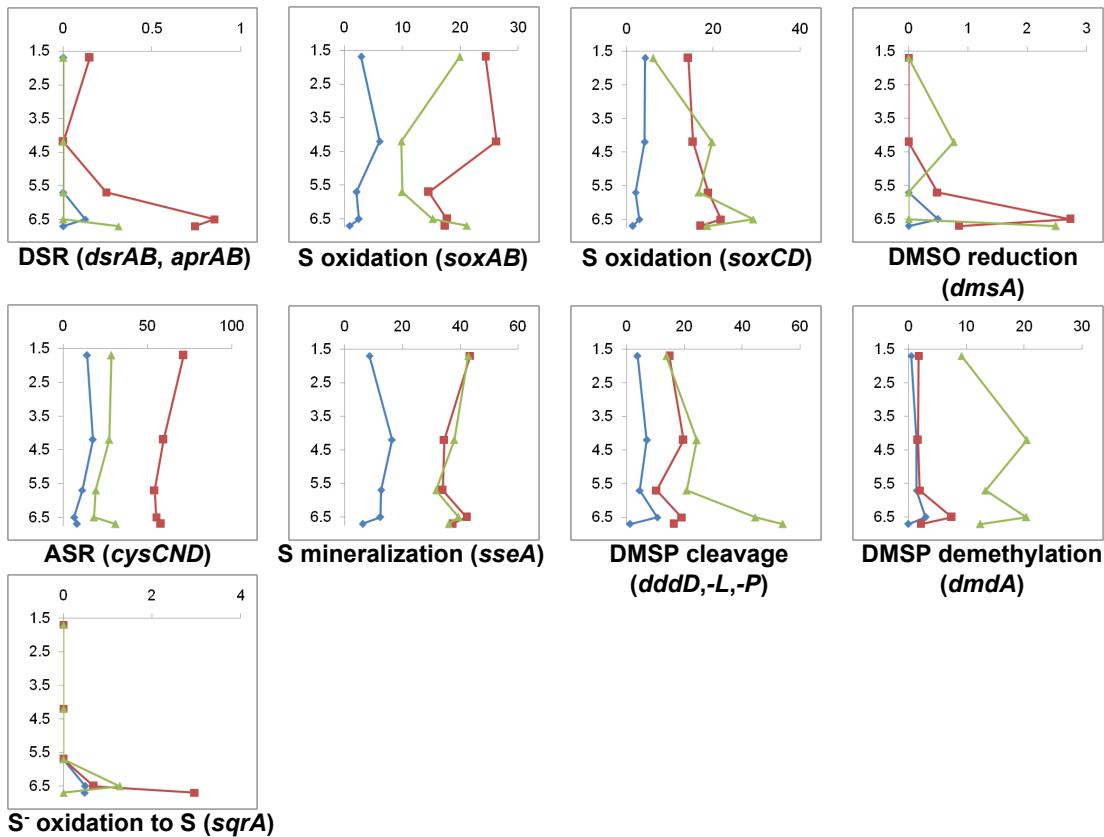


Figure 4.13: Vertical profiles of potential for sulphur conversions for each size fraction in Organic Lake. The y-axis shows sample depths (m) and the x-axis shows counts of marker genes normalised to 100 Mbp of DNA sequence. The 0.1, 0.8, 3.0 μm size fractions are shown as blue, red and green, respectively. Counts for marker genes for the same pathway or enzyme complex were averaged and those from different pathways were summed. For marker gene descriptions see Table 4.2 and Table 4.3.

also made for Ace Lake, although in this system the presence of a dense layer of green sulphur bacteria with the potential to fix nitrogen augments the nitrogen cycle (Lauro *et al.*, 2011).

4.4.7 Molecular basis for unusual sulphur chemistry

Several meromictic hypersaline lakes in the Vestfold Hills, including Organic Lake, with practical salinity >150 are characterized by an absence of hydrogen sulphide and photoautotrophic sulphur bacteria (Burke and Burton, 1988). Although sulphate is present (Franzmann *et al.*, 1987b), geochemical conditions of these lakes are not conducive to dissimilatory sulphur cycling between sulphur oxidising and sulphate reducing bacteria typical of other stratified systems such as Ace Lake (Ng *et al.*, 2010; Lauro *et al.*, 2011). Consistent with this, potential for dissimilatory sulphate reduction represented by dissimilatory sulfite reductase (*dsrAB*) and adenylylsulphate reductase (*aprAB*) linked to sulphate-reducing *Deltaproteobacteria* (Table 4.8) was low in Organic Lake. Sulphate-reduction potential was confined to the 6.7 m sample (Figure 4.13) where oxygen concentration was lowest and *Deltaproteobacteria* were present (Figure 4.6).

Capacity for oxidation of reduced sulphur compounds, represented by the sulphur oxidation multienzyme genes (*soxAB*), was present throughout the water column (Figure 4.13) and linked primarily to *Alpha-* and *Gammaproteobacteria* (Table 4.8). Sulphur-oxidising *Alpha-* and *Gammaproteobacteria* are known to oxidise sulphur compounds, such as thiosulphate, aerobically. Although a small proportion of *Gammaproteobacteria* had the capacity for autotrophy (see 4.4.5), the majority of sulphur-oxidizers were likely chemolithoheterotrophs as they were related to heterotrophic *Marinobacter* and *Roseobacter* clade. The sulphur dehydrogenase genes *soxCD* linked to *Alpha-* and *Gammaproteobacteria* were similarly present throughout the water column. *soxCD* are accessory components of the Sox enzyme system without which complete oxidation of thiosulphate cannot occur (Friedrich *et al.*, 2005). Thus the presence of *soxCD* indicates complete oxidation likely occurs, although the different distribution of *soxAB* and *soxCD* in the water column (Figure 4.13) suggests a proportion of the community may lack *soxCD* and deposit sulphur.

Sulphur-oxidising *Epsilonproteobacteria* possessing *soxAB* genes (Table 4.8) were present only in the deep zone of Organic Lake (Figure 4.6) and were related to autotrophic deep sea sulphur-oxidisers, some members of which are capable of anaerobic sulphur oxidation using nitrate (Yamamoto and Takai, 2011). It is unlikely that appreciable sulphur oxidation occurs in the deep zone as the known terminal electron acceptors, oxygen and nitrate, are depleted and the abundance of sulphur oxidising *Epsilonproteobacteria* is low (Figure 4.6). *Epsilonproteobacteria* were also linked to a capacity for oxidation of sulphide to elemental sulphur by utilising sulphide:quinone oxidoreductase (*sqrA*) (Figure 4.13, Table 4.8). In this pathway, sulphur is released as polysulphides, which is a potential biological source of the abundant polysulphides that have been detected in the lake (Roberts *et al.*, 1993).

It is likely that the limited anaerobic dissimilatory sulphur cycle contributes to the accumulation of DMS in Organic Lake in the deep zone. In the upper mixed zone, DMS could potentially be oxidized as a carbon and energy source or utilized as an electron donor by sulphur-oxidising bacteria (Schäfer *et al.*, 2010). In anoxic zones, methanogenic *Archaea* or sulphate-reducing bacteria are the main organisms known to breakdown DMS (Schäfer *et al.*, 2010). Methanogens and genes involved in methanogenesis were not detected, nor has methane been detected (Gibson *et al.*, 1994) leaving sulphate-reduction the most likely route of DMS catabolism. The low dissimilatory sulphate reduction potential in the deep zone coupled with the relatively stagnant waters would likely minimize DMS oxidation and loss by ventilation. DMS would therefore be expected to accumulate in the deep zone if production rates were higher than breakdown.

To determine the source of high DMS in the bottom waters of Organic Lake, the genes involved in DMS formation were surveyed. Genes for DMSP lyases *dddD*, *dddL* and *dddP*, were detected in Organic Lake at levels comparable to other dominant processes such as respiration and fermentation (Figure 4.13) indicating DMSP is an important carbon and energy source in Organic Lake. *dddD* was the most abundant of

Table 4.8: Contribution of different taxonomic groups to counts of marker genes involved in sulphur conversions. The taxon with the largest contribution to each process is highlighted in blue.

| Taxon | DSR | <i>S</i> oxidation <i>sqrA</i> | <i>S</i> assimilation | <i>S</i> mineralisation | DMSO reduction |
|------------------------------|------|-----------------------------------|-----------------------|-------------------------|----------------|
| <i>Acidobacteria</i> | 0 | 0 | 0.03 | 1.03 | 0 |
| <i>Actinobacteria</i> | 0 | 0 | 0.15 | 0 | 0 |
| <i>Alphaproteobacteria</i> | 0 | 2.05 | 0 | 0.85 | 11.7 |
| <i>Aquificae</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Bacteroidetes</i> | 0 | 0 | 5.06 | 0.20 | 0 |
| <i>Betaproteobacteria</i> | 0 | 1.09 | 0 | 2.07 | 0.69 |
| <i>Chlorobi</i> | 0 | 0 | 0 | 0.03 | 0.15 |
| <i>Chloroflexi</i> | 0 | 0 | 0 | 0.30 | 0 |
| <i>Chrysigenetes</i> | 0 | 0 | 0 | 0.01 | 0 |
| <i>Cyanobacteria</i> | 0 | 0 | 0 | 0.13 | 0.05 |
| <i>Deferribacteres</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Deinococcus-Thermus</i> | 0 | 0 | 0 | 0.09 | 0 |
| <i>Delta proteobacteria</i> | 0.19 | 0 | 0 | 0.56 | 0.20 |
| <i>Epsilonproteobacteria</i> | 0 | 0.03 | 0.39 | 0.13 | 0 |
| <i>Firmicutes</i> | 0 | 0 | 0 | 0.22 | 0.09 |
| <i>Fornicata</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Fusobacteria</i> | 0 | 0 | 0 | 0.03 | 0.13 |
| <i>Gammaproteobacteria</i> | 0 | 2.64 | 0 | 22.4 | 14.0 |
| <i>Nitrospirae</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Planctomycetes</i> | 0 | 0 | 0 | 0.03 | 0 |
| <i>Spirochaetes</i> | 0 | 0 | 0 | 0 | 0.03 |
| <i>Thermobaculum</i> | 0 | 0 | 0 | 0 | 0.08 |
| <i>Thermotogae</i> | 0 | 0 | 0 | 0.01 | 0 |
| <i>Verrucomicrobia</i> | 0 | 0 | 0 | 0.18 | 0 |
| <i>Crenarchaeota</i> | 0.02 | 0 | 0 | 0 | 0 |
| <i>Euryarchaeota</i> | 0 | 0 | 0 | 0.02 | 0.12 |
| <i>Alveolata</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Euglenozoa</i> | 0 | 0 | 0 | 0 | 0.03 |
| <i>Opistokonta</i> | 0 | 0 | 0 | 0.11 | 0.03 |
| <i>Rhodophyta</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Strameopiles</i> | 0 | 0 | 0 | 0 | 0 |
| <i>Viridiplantae</i> | 0 | 0 | 0 | 0.03 | 0.15 |

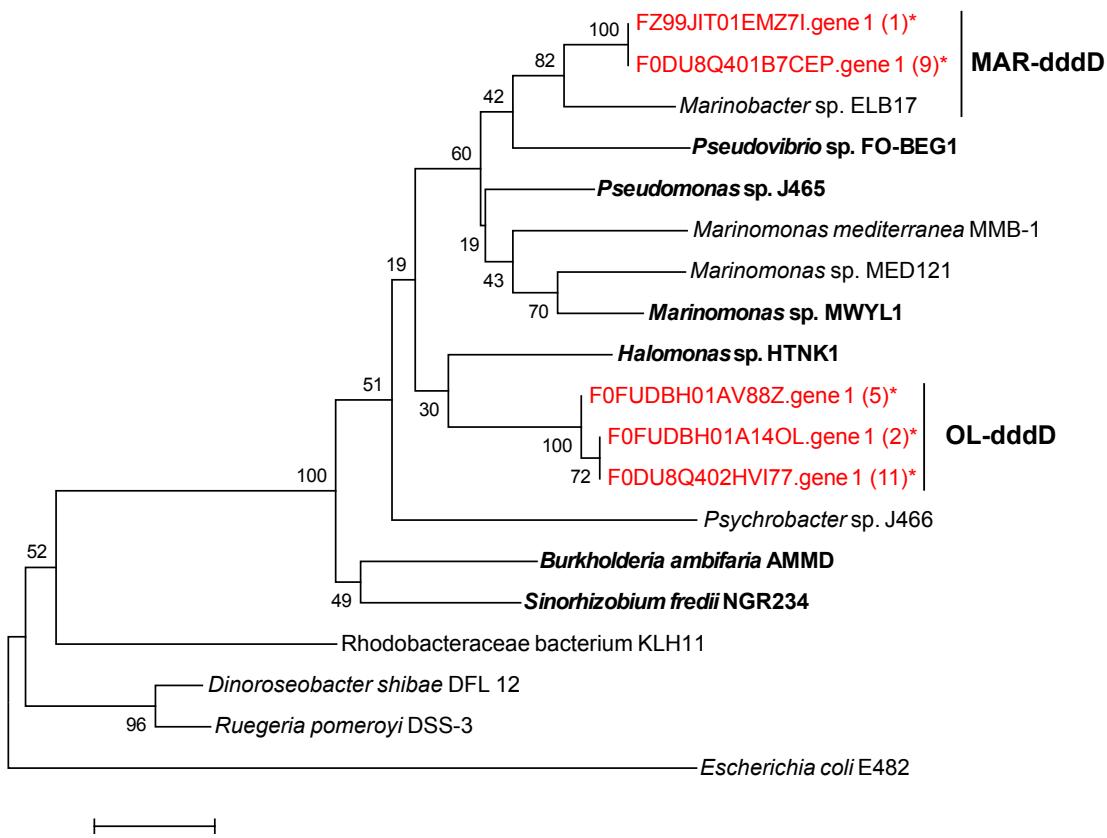


Figure 4.14: Phylogeny of DddD DMSP lyase homologues. *E. coli* carnitine coenzyme A transferase was used as an outgroup. *Dinoroseobacter shibae* DFL 12 and *Ruegeria pomeroyi* DSS-3 homologues are a non-functional outgroup (Todd *et al.*, 2011). The tree was computed from a 75 amino acid region within the conserved amino-terminal class III coenzyme A domain (CaiB) using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP lyase activity are shown in bold. Accession numbers from top to bottom are: EBA01716, AEV37420, ACY01992, ADZ91595, EAQ63474, ABR72937, ACV84065, ACY02894, ABI89851, YP_002822700, EEE36156, ABV95365, AAV94987 and EGB36199.

the Organic Lake DMSP lyases (Table 4.6) and comprised two main types: MAR-dddD and OL-dddD (Figure 4.14). Neither of these types clustered with the non-functional *Dinoroseobacter shibae* DFL 12 and *Ruegeria pomeroyi* DSS-3 *dddD* homologues (Todd *et al.*, 2011) or carnitine coenzyme A transferase outgroups, thereby providing support for their proposed role as functional DMSP lyases. The MAR-dddD type includes the *Marinobacter* sp. ELB17 *dddD* homologue, and MAR-dddD sequences were most abundant on the 0.8 µm fraction where *Marinobacter* OTUs were also abundant, indicating MAR-dddD derives from Organic Lake *Marinobacter* (Figure 4.14). OLn-dddD did not have a close relative from cultured bacteria making its precise taxonomic origins uncertain. The abundance of OL-dddD on the 3.0 µm fraction suggests it originates from *Alphaproteobacteria*. OL-dddD containing contigs carried genes of mixed *Alpha-* and *Gammaproteobacterial* origin supporting its provenance from one of these classes and

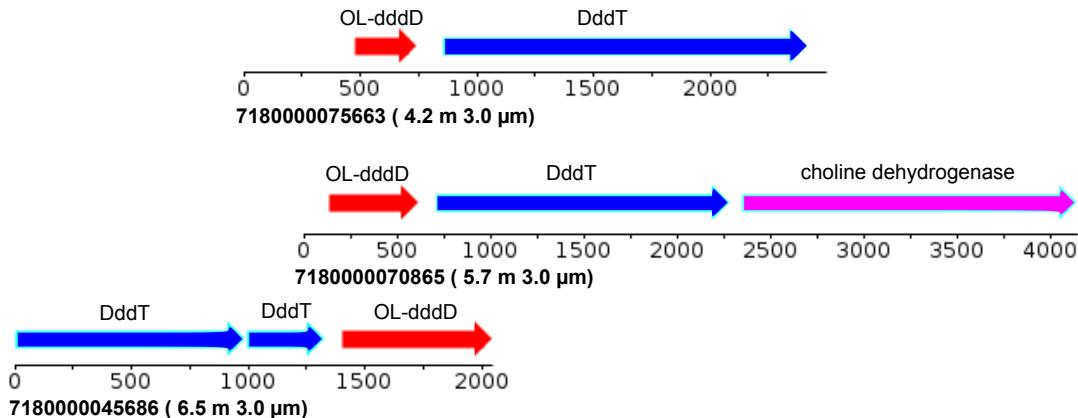


Figure 4.15: Genomic maps of Organic Lake scaffolds containing the OL-dddD homologue. DddT and choline dehydrogenase had best BLAST matches to *Halomonas* sp. HTNK1 (*Gammaproteobacteria*) and *Hoeflea phototrophica* DFL-43 (*Alphaproteobacteria*), respectively. The numbers represent base pairs. The sample depth and filter from which the scaffold was assembled is shown in parentheses beside the scaffold ID.

consistent with the “pick n’ mix” arrangement of genes found beside sequenced *dddD* regions (Johnston *et al.*, 2008) (Figure 4.15). Adjacent to OL-dddD was *dddT* (Figure 4.15), a betaine, choline, carnitine transporter (BCCT) family protein that likely functions in substrate import, demonstrating OL-dddD forms an operon-like structure, similar to *Halomonas* sp. HTNK1 (Todd *et al.*, 2010).

Two *dddL* groups were detected in Organic Lake: SUL-dddL and MAR-dddL (Figure 4.16). The former includes the *Sulfitobacter* sp. EE-36 *dddL* and the latter the *Mari nobacter manganoxydans* MnI7-9 homologue indicating they originate from *Roseobacter* clade and *Gammaproteobacteria*, respectively. *Sulfitobacter* sp. EE-36 has demonstrated DMSP lyase activity and the *dddL* gene alone is sufficient for DMS generation (Curson *et al.*, 2008). These data indicate that the Organic Lake members of the SUL-dddL group perform the same functional role. The MAR-dddL clade appears to be an uncharacterized branch of the *dddL* family. emphdddP was detected as the least abundant of the DMSP lyases (Table 4.6). Phylogenetic analyses showed Organic Lake *dddP* likely originate from *Roseovarius* (Figure 4.17). The Organic Lake sequences formed a clade with the functionally verified *Roseovarius nibinhicensis* ISM *dddP* (Todd *et al.*, 2009).

A single type of DMSP demethylase, *dmdA* was identified. It clustered with *Roseobacter* clade *dmdA* (Figure 4.18), corresponding to the marine clade A (Howard *et al.*, 2006), and includes the functionally verified *R. pomeroyi* DSS-3 homologue. These data indicate that the Organic Lake sequences correspond to true DMSP demethylases and not related glycine cleavage T proteins or aminomethyltransferases (Howard *et al.*, 2006). DMSP cleavage appears to be a significant source of DMS in Organic Lake. DMSP likely originates from *Bacillariophyta* or *Dinoflagellida* as Organic Lake *Dunaliella* have been reported not to produce DMSP in culture (Franzmann *et al.*, 1987b). Based on the abundance of marker genes, DMSP cleavage is predicted to occur at highest levels in the deep zone (Figure 4.13) where the DMS concentration has been

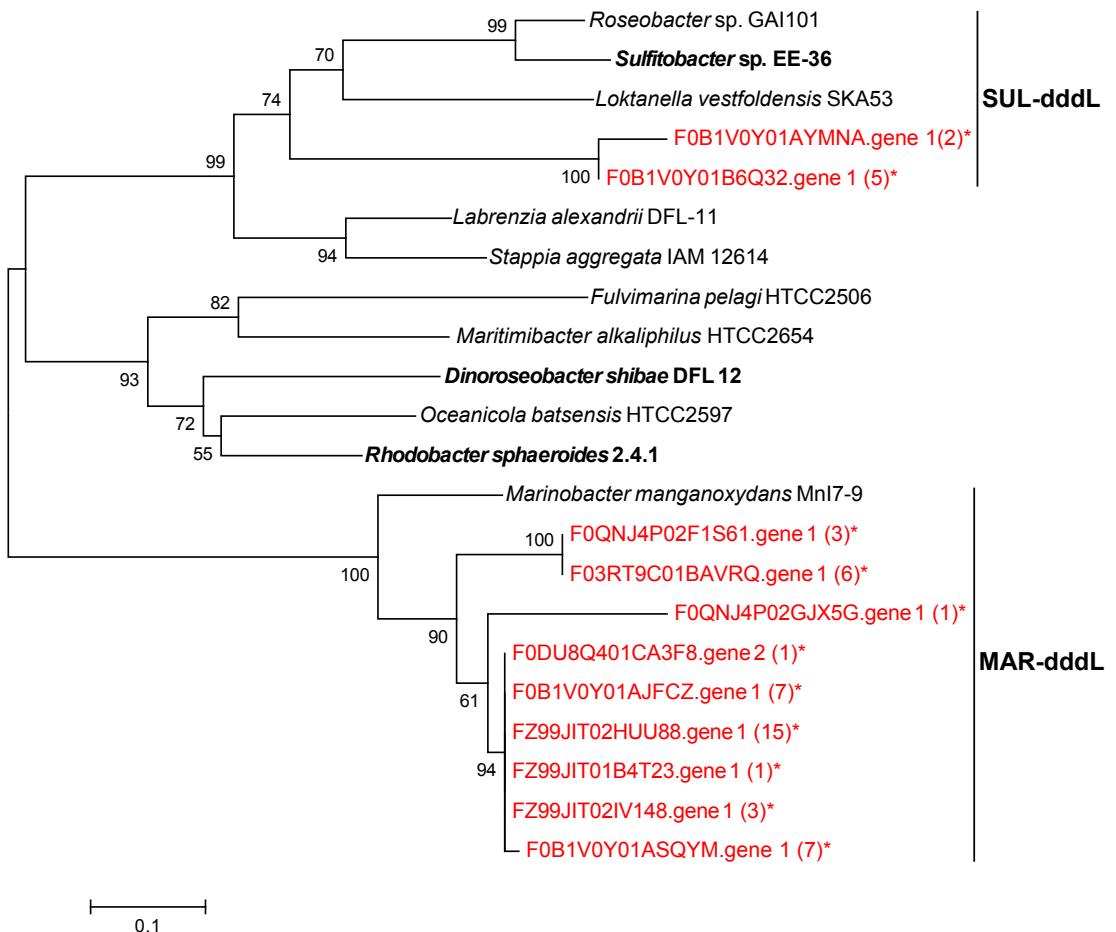


Figure 4.16: Phylogeny of DddL DMSP lyase homologues. The tree was computed from an 84 amino acid N-terminal region using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP lyase activity are shown in bold. Accession numbers from top to bottom are: EEB86351, ADK55772, EAQ07081, EEE47811, EAV43167, EAU41122, EAQ10619, ABV95046, EAQ04071, ABA77574 and EHJ04839.

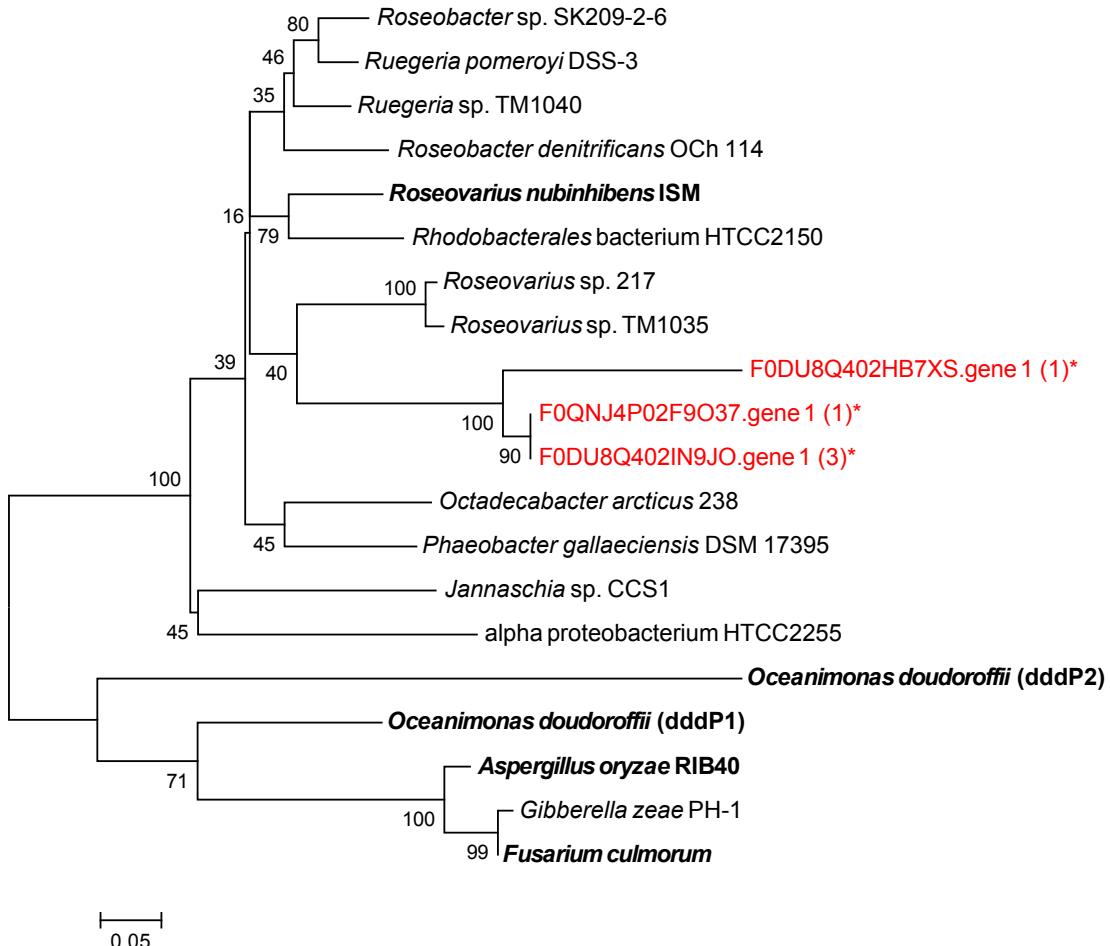


Figure 4.17: Phylogeny of DddP DMSP demethylase homologues. The tree was computed from a 129 amino acid C-terminal region using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP lyase activity are shown in bold. Accession numbers from top to bottom are: ZP_01755203, YP_167522, YP_613011, YP_682809, EAP77700, ZP_01741265, ZP_01036399, ZP_01881042, ZP_05063825, AFO91571, YP_509721, ZP_01448542, AEQ39103, AEQ39091, XP_001823911, XP_389272 and ACF19795.

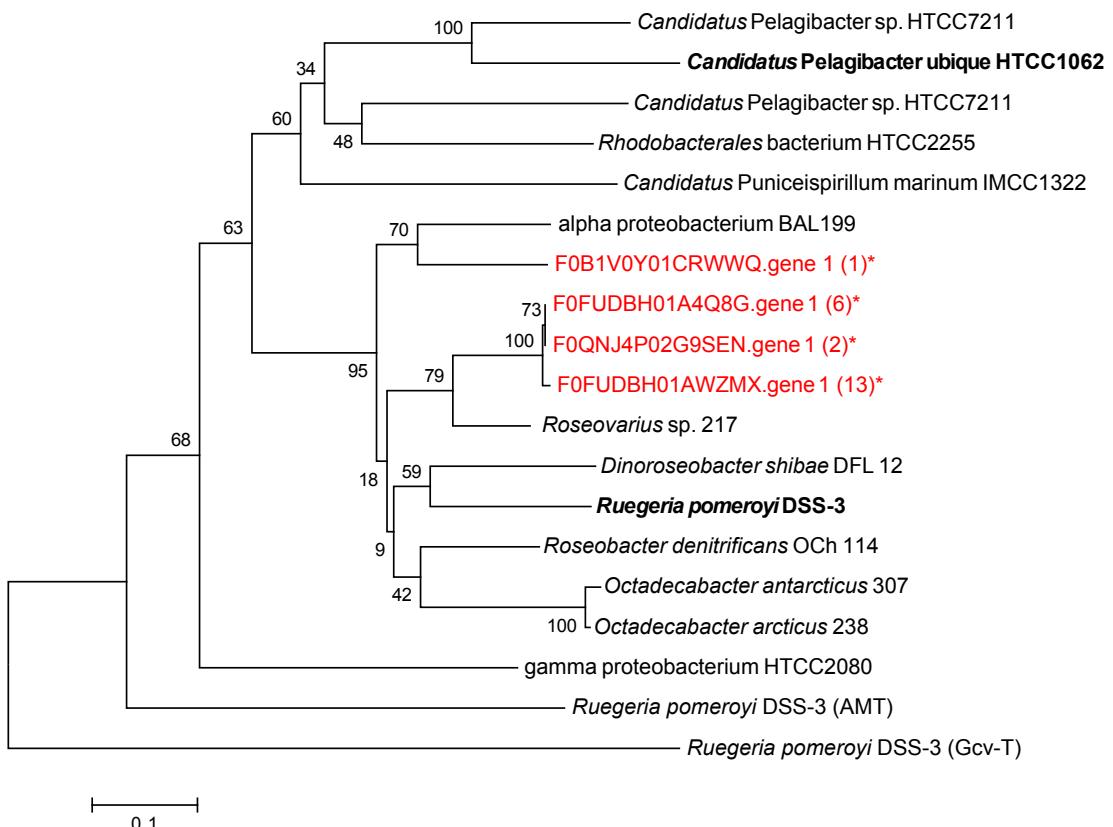


Figure 4.18: Phylogeny tree of DmdA DMSP demethylase homologues. The tree was computed from a 128 amino acid region using the neighbour-joining algorithm. Organic Lake sequences from this study are shown in red and marked with an asterisk (*). Numbers in parentheses are counts of sequences that clustered with the Organic Lake homologue shown in the tree with 90% amino acid identity. Sequences with confirmed DMSP demethylase activity are shown in bold. Accession numbers from top to bottom are: EDZ60447, YP_265671, EDZ61098, EAU51039, YP_003550401, EDP61332, EAQ26389, ABV94056, AAV94935, AAV95190, EDY79173, EDY89914, EAW42451, AAV94935 and AAV97197.

measured to be highest (Deprez *et al.*, 1986; Franzmann *et al.*, 1987b; Gibson *et al.*, 1991; Roberts and Burton, 1993; Roberts *et al.*, 1993). DMS can also be produced in anoxic environments from the reduction of DMSO, degradation of sulphur containing amino acids, and sulphide methylation (Schäfer *et al.*, 2010). Our data indicate that some DMSO reduction linked to *Firmicutes* could occur, but is not likely a major pathway (Figure 4.13), and the potential for the other DMS yielding processes could not be determined because the enzymes involved in these pathways have not been established. When cultivated, *Halomonas* isolates from Organic Lake produced DMS from cysteine (Franzmann *et al.*, 1987b) providing some evidence that DMS production from anaerobic degradation of amino acids can occur. Abiotic pathways for anaerobic production of DMS have also been proposed (Roberts *et al.*, 1993).

The potential for DMSP cleavage was more than twice that of DMSP demethylation (Figure 4.13). This is unusual compared to the marine environment or Ace Lake where DMSP demethylation potential is much higher than cleavage (Table 4.6). Previous estimates have similarly shown marine environments to have demethylation potential up to two orders of magnitude higher than cleavage (Howard *et al.*, 2008; Todd *et al.*, 2009, 2011; Reisch *et al.*, 2011). The frequency of DMSP lyase genes *dddD* and *dddL* in Organic Lake exceeded those of all other environments, except Punta Cormorant hypersaline lagoon, where *dddL* abundance was comparable (Table 4.6). This suggests selection in Organic Lake for DMSP cleavage due to functional advantage and/or selection for taxa that carry DMSP lyase genes. There is evidence that high DMSP cleavage potential is adaptive in hypersaline systems, as a high proportion of *ddd* genes were similarly detected in Punta Cormorant hypersaline lagoon and saltern ponds (Raina *et al.*, 2010). Determination of the taxonomic composition of these other hypersaline environments could indicate whether selection is occurring for functional capacity or on a taxonomic level if the taxonomic composition between these systems was significantly different but abundance of DMSP lyase genes were high.

The accumulated DMS in Organic Lake suggests conditions in Organic Lake favor the relatively inefficient lysis pathway, where both sulphur and carbon is lost to the organism performing the DMSP lysis, over the more ‘thrifty’ demethylation pathway. This is particularly pertinent to the *Roseobacter* lineages that can also perform either process. One possibility that has been proposed is that when sulphur is in excess and the organism can easily assimilate alternative sulphur sources, the lysis pathway may be competitive (Johnston *et al.*, 2008). This may be particularly the case in hypersaline systems if higher concentrations of DMSP are being produced as an osmolyte.

4.5 Conclusions

Through the use of shotgun metagenomics and size partitioning of samples, we discovered that the Organic Lake system is dominated by heterotrophic bacteria related to *Psychroflexus*, *Marinobacter* and *Roseovarius* with primary production provided largely by chlorophyte algae related to *Dunaliella*. Genetic potential for oxidation of fixed carbon by heterotrophic bacteria occurs greatly in excess of carbon fixation, suggesting

possible net carbon loss. However, by linking key metabolic processes to the dominant heterotrophic lineages we uncovered processes that were unusually abundant in Organic Lake that may serve to maximize exploitation of limited resources and minimize loss. Recalcitrant polymeric algal material and particulate matter is likely remineralized by *Psychroflexus* in the upper mixed zone and by *Firmicutes* in the deep zone to provide labile substrates for use by other heterotrophic bacteria. The generalist *Marinobacter* and *Roseovarius* lineages were associated with abundant genes involved in rhodopsin-mediated and AAnP photoheterotrophy; the latter of which was more abundant in Organic Lake than any other system surveyed. Potential for chemolithoheterotrophy, sulphur oxidation and CO oxidation was also high, and along with photoheterotropy, may provide a supplementary energy source if organic carbon becomes limiting.

In addition to being able to describe the functional capacities and potential importance of poorly understood microbial processes occurring in the lake (e.g. photoheterotrophy by *Alphaproteobacteria*), we were able to answer targeted questions about the biology of the unusual lake sulphur chemistry. The low potential for dissimilatory sulphur cycling in the deep zone and relatively stable waters, combined with the generation of DMS from DMSP, facilitate the accumulation of a high level of DMS in the lake. It appears *Marinobacter* and *Roseovarius* play a key role in DMS formation by cleaving DMSP generated by upper mixed zone eucaryal algae. The remarkable abundance of DMSP lyase genes suggests DMSP is a significant carbon source in Organic Lake and the cleavage pathway provides a selective advantage under the unique constraints of the Organic Lake environment.

In view of the minimal capacity for biological fixation of carbon and nitrogen, and yet organic richness, including high levels of DMS, in Organic Lake, we evaluated what input the lake may have received throughout its relatively brief ~3,000 year history. The volume of the lake is small ($\sim 6 \times 10^4$ m³), and exogenous input may occur from guano deposits in a small penguin rookery nearby the lake, through giant petrel or skua predation and defecation, and/or by decaying animal carcasses such as elephant seals which can weigh on the order of 1 ton and are present near the lake. It is also possible that during isolation from the ocean, the base of the water column in the marine basin that formed the lake may have acted as a sump for organic material. Phytoplankton blooms and benthic mats tend to make coastal marine basins very productive, and organic matter that sediments out of the surface waters will become trapped in the denser, more saline bottom layers (Bird *et al.*, 1991). Retention of captured organic matter in the lake may also have been facilitated by Organic Lake having become highly saline quickly (Bird *et al.*, 1991). Studies in the future that experimentally determine exogenous input and historical lake dynamics (e.g. stable isotope and biomarker analyses of lake sediment), the role of benthic communities, and metaproteogenomic analyses of interannual community composition and function, will provide improved knowledge of the unusual biogeochemistry of Organic Lake and better enable predictions to be made about how the lake may be affected by ecosystem changes.

Chapter 5

General discussion, future work and conclusions

Metagenomics has proven to be an effective way to map the diversity of Antarctic lake ecosystems and provide hints of how they work. In combination with metaproteomics and abiotic parameters, in-depth descriptions were achieved of the ecosystem functions of Ace Lake and Organic Lake. The most noteworthy contribution of these studies has been to describe taxa and microbial processes previously unknown in these lakes, and from these descriptions, generate testable hypotheses of population and ecosystem level function.

5.1 Possible future work on Organic Lake

To establish a picture of the biotic composition and function of Antarctic lake communities, an observation-driven approach was utilised that allowed for completely new discoveries about these systems to be made. Bioinformatic pipelines and theoretical models to do this have now been established. These can be applied to future, similarly observation-based studies of Ace and Organic Lakes to define how they change over time and test our existing models of how the lakes function. It is also clear that a systems level understanding of the lakes is bolstered by having well-characterised isolates related to members in the community. Guided by the studies from this thesis, isolation and characterisation of key members of the community, such as *Pyramimonas* spp., Organic Lake phycodnavirus (OLPV), Organic Lake virophage (OLV), *Marinobacter* spp. and *Roseovarius* spp. from Organic Lake, could be attempted in the future. Other complementary techniques, such as single-cell and single-virus genomics and stable isotope probing (SIP), hold promise for learning about specific populations in the community. Some experiments that could be used to test specific hypotheses generated by the study of Organic Lake are discussed below, along with their expected outcomes and implications.

5.1.1 Organic Lake community dynamics

To establish a complete understanding of Organic Lake requires descriptions of the microbial community over different time scales. Certain bacterial taxa have been observed to peak in abundance in Organic Lake during summer indicating seasonal fluctuations exist (James *et al.*, 1994). The Organic Lake community profiles from the summer of 2006 and 2008 varied substantially (chapters 3 and 4) identifying further community members that change with time. However, the rates of change and degree of variance for these populations are unknown. Currently, a baseline of the microbial community diversity over an annual cycle has not been established and is the next step required for improved modelling of the Organic Lake ecosystem.

Metagenomic and functional data obtained over the course of a year can determine how the microbial community responds to the large seasonal changes in temperature, light and ice-cover. Winter samples in particular can establish how the community persists when the lack of light is expected to curtail photosynthetic production. Paired metagenomic and metaproteomic analyses of the coastal Southern Ocean from winter and summer found winter samples were dominated by active chemolithoautotrophs including sulphur-oxidising *Gammaproteobacteria* and ammonia oxidising *Crenarchaeota* (Grzymski *et al.*, 2012; Williams *et al.*, 2012). If the Organic Lake ecosystem reflects that of the coastal Southern Ocean, the small population of chemolithoautotrophic sulphur-oxidising *Proteobacteria* found to be present (chapter 4) are expected to become dominant in the winter. Metagenomic analysis indicated recycling of reduced nitrogen compounds predominates in Organic Lake and a lack of capacity to form oxidised nitrogen compounds (chapter 4). If this model of nitrogen cycle holds true over the year, we can expect ammonia oxidising microorganisms will not play a part in the community as they do in other similar environments (Voytek *et al.*, 1999; Grzymski *et al.*, 2012; Williams *et al.*, 2012).

Photosynthetic nanoflagellates detected in Organic Lake, such as *Pyramimonas*, are capable of mixotrophy (Bell and Laybourn-Parry, 2003) and likely persist over the winter by switching to heterotrophy. Strategies that might also be employed by other photosynthetic algae include utilising starch reserves or developing cyst forms. In Lake Bonney, trends between RuBisCO expression and irradiance levels differed with RuBisCO type and depth suggesting adaptations to the polar night varies between species (Kong *et al.*, 2012b). Phototrophic eucaryotes in Organic Lake likely have similarly diverse strategies to persist during prolonged darkness. Combined metagenomic and metaproteomic analysis can determine which phototrophic eucaryotes are found during the winter and indicate what metabolic processes they employ in the absence of light.

Some members of the Organic Lake population appear to be persistent over long time scales. For example, *Dunaliella* (Franzmann *et al.*, 1987b), choanoflagellates (van den Hoff and Franzmann, 1986), *Marinobacter*, *Psychroflexus Roseovarius* and *Halomonas* (Bowman *et al.*, 2000a) have been recorded in Organic Lake previously and were detected again in studies from this thesis (chapters 3 and 4). How they vary

throughout the year and between years is unclear. Organic Lake has been shown to be physically variable over a decadal time scale with changes in water level of \sim 1 m leading to large changes in the water column structure (Gibson *et al.*, 1995; Gibson and Burton, 1996). Notably, the pycnocline in Organic Lake occurred at 5.7 m in 2008; much lower than 3.5 m where it first reported in 1978 (Franzmann *et al.*, 1987b). If the lake completely mixes, this would challenge the oxygen-sensitive microbes and processes occurring in the deep zone. In West Antarctic lakes physical changes have been correlated with large and rapid ecological responses (Quayle *et al.*, 2002). Substantial differences were observed in *cbbM* gene copy number and vertical distribution between three field seasons in Lake Bonney (Kong *et al.*, 2012a) indicating sensitivity to environmental factors applies to permanently ice-covered lakes and is a general property of Antarctic lakes. However, the amount of seasonal variability needs to be established in Organic Lake for differences over longer time scales to be determined.

Sampling over the course of a summer season would be particularly valuable as these samples can be compared to the available summer datasets to give a better indication of the degree of variability between summer seasons. There was a change in the viral composition and abundance between November and December 2008 samples when the lake began to thaw (chapter 3). Similarly, there were differences in the eukaryotic community when the lake was completely thawed in 2006 (chapter 3) and the profile in November 2008 when the lake was ice-covered (chapter 4). It is unclear whether the variation between the summer samples, December 2006, November 2008 and December 2008 are due to differences in ice-cover/light, sampling locations (2006 and December 2008 samples were littoral, while November 2008 was over the deepest point of the lake) or other environmental factors. Metagenomic sequencing of a summer time course would provide a high resolution view of changes in the microbial community over the season that can be related to environmental conditions. This time course can be compared back to the summer 2006 and 2008 samples indicating if the differences that have been observed fall within normal range of variability over the summer season and also help interpret what factors were driving those variations.

Metagenomic sequencing of samples from a summer time course can also validate the OLV–OLPV–host model formulated in chapter 3. Genomic analysis indicated OLV is a new member of the virophage virus family, which uses OLPV as a ‘helper’ virus to complete its replication cycle but impairs its helper’s infectivity in the process. The likely host of OLV and OLPV was determined to be the unicellular alga, *Pyramimonas*. The inferred interaction of the OLV, OLPV and host was used to generate a Lotka–Volterra model of their population dynamics that showed if OLV acts as a ‘predator of a predator’, this would lead to increased frequency of population cycles. Observations of changes in OLV, OLPV and *Pyramimonas* abundances over a summer time course can establish if a relation exists between them. An observable trend between these populations would lend further support to the claim that OLV–OLPV and *Pyramimonas* are a virophage–virus–host system. In the model, the density of each population oscillates in a phase-shifted manner in the sequence: prey, predator and predator of predator.

Observed population changes can determine if the data fits the model, suggest if additional modifications to the existing model need be made or if alternative models would better fit the data. If the dynamics fit the model, observed population densities can be used to derive the parameters that govern their interaction.

5.1.2 OLV physiology and ecology

Since the discovery of the OLV, two other members of the virophage family have been reported that have afforded a new perspective on OLV. The first of these is the Mavirus (for Maverick virus) so named for its evolutionary relationship with the Maverick/Politon class of eucaryotic transposons (Fischer and Suttle, 2011). Like Sputnik, Mavirus has an absolute requirement for a helper virus, *Cafteria roenbergensis* virus (CroV), to replicate and is deleterious to its helper (Fischer and Suttle, 2011). The other virophage reported was called Sputnik 2 (its genome is almost identical to Sputnik), but unlike Sputnik, it was found both as a separate genome and integrated in its helper lentille virus (Desnues *et al.*, 2012). These two virophage systems in culture are associated with heterophic protists making OLV the only genomic sequence currently available for virophage affecting cosmopolitan phytoplankton species. Therefore isolating OLV, acquiring genomes of OLV relatives in Antarctic lakes and determining fundamental properties of OLV physiology and dynamics would contribute immensely to our understanding of the evolution and diversity of virophages in general and is highly relevant to other aquatic systems.

The evidence that Mavirus has the same ‘virophage’ phenotype as Sputnik strengthens the inference that OLV does as well. This is in part because Mavirus and CroV are quite divergent from Sputnik and mimivirus respectively, yet retain the same traits (Fischer *et al.*, 2010; Fischer and Suttle, 2011), suggesting a common feature of the whole lineage. Moreover, as both mimivirus and CroV encode much of their transcriptional machinery, they do not localise to the nucleus during infection but generate the viral factory in the cytoplasm of their host (La Scola *et al.*, 2008; Fischer *et al.*, 2010). Sputnik and Mavirus appear to replicate in the giant virus factory making use of helper virus’ replication and transcription systems, not the host cell’s (La Scola *et al.*, 2008; Claverie and Abergel, 2009; Fischer and Suttle, 2011). In this way, the replicative strategy of the mimivirus lineage makes them vulnerable to virophages parasitising necessary resources during replication, which is a likely cause of reduced helper virus production (Claverie and Abergel, 2009; Fischer and Suttle, 2011; Fischer, 2011). As OLPV encodes all RNA polymerase subunits (see GenBank accession HQ704802), this is evidence that it replicates solely in the cytoplasm, thereby making it susceptible to a virophage. A larger census of virophage and helper viruses would be able to show if virophages only associate with members of the nucleo-cytoplasmic large DNA virus (NCLDV) clade that replicate in the cytoplasm. Ultimately, definitive confirmation of a detrimental effect on the helper by OLV co-infection is likely only possible by observing infection in culture.

Although Sputnik and Mavirus have similar effects on their helper viruses, their

infection strategies appear different raising interesting considerations for OLV. Mavirus can be independently phagocytosed by *Cafeteria roenbergensis* in the absence of CroV – perhaps serving as a defence for the host in the case of CroV infection (Fischer and Suttle, 2011). In support of this, Mavirus possesses a retroviral-family integrase that is theorised to have been separately acquired as a way to stabilise the relationship between the ancestral Mavirus and its cellular host (Fischer and Suttle, 2011) although as yet, Mavirus has not been reported to be integrated in *Cafeteria roenbergensis*. In contrast, Sputnik seems to associate directly with the fibrils that coat the mimivirus virion, not the host *Acanthamoeba* cell (Boyer *et al.*, 2011). These two modes of infection, that is, a virophage-bearing host being infected by a giant virus *vs.* a host being infected by a virophage-bearing giant virus, produces different selection pressures that would lead to distinct effects on the population dynamics in the environment. Detection of complete OLV genomes in the 0.8–0.1 μm size fraction indicates it was captured in association with larger particles but does not distinguish how it is transmitted.

Determination of the mode of infection for OLV would improve predictive modelling of OLV driven impacts on algal blooms. This could be achieved by observation of OLV-OLPV infections in culture. However, recently developed methods for flow cytometric sorting to obtain single-cell amplified genomes (SAGs) or single virus genomes (SVGs) (Martínez Martínez *et al.*, 2011; Allen *et al.*, 2011) could be applied to Organic Lake water samples to study the mode of OLV, OLPV and host interaction. For example, a sample population of single host algal cells could be fluorescence sorted and screened for the presence of OLPV and OLV. At the same time, a sample population of single OLPV particles could be sorted and screened for the presence of OLV. This could reveal if OLV is able to associate with the host independently of OLPV or *vice versa*. Examination of the infected algal cells could also establish fundamental properties of the algae and virus populations such as proportions of infected cells and proportions of OLPV infections that include an OLV. As these methods are quite new, an experiment of this kind would require optimisation to successfully capture an adequate sample of infected cells and target giant viruses. However, just obtaining SVGs would provide invaluable information on virus diversity and genetic content making it worthwhile pursuing as a complement to metagenomic sequencing.

A final consideration for OLV and OLPV dynamics is suggested by the discovery that Sputnik 2 can integrate into its helper lentille virus (Desnues *et al.*, 2012). Integration of Sputnik 2 is localised to a 352 bp region corresponding to the Sputnik V6 gene that encodes a collagen-like repeat-containing protein shared by Sputnik 2, lentille virus and mamavirus (Desnues *et al.*, 2012). OLV and OLPV share a homologous collagen-like repeat-containing protein that may similarly function as a site of integration. As yet, the conditions and mechanism by which Sputnik 2 integrates is unknown. Screening of metagenomic assemblies, SVGs or isolated OLPV genomes for integrated virophages can determine if this occurs in the environment. Integration of OLV and OLPV could have interesting evolutionary functions. One possibility is that integration of virophages may function analogously to lysogeny in bacteriophages

if provirophages are inactive. In this scenario, virophages would integrate into their helper virus when helper virus densities are low to avoid driving their helpers to extinction. On the other hand, if integration does not entail dormancy of the virophage, it could function as a means to ensure transmission. In either case, evidence of integration can be found in future genomic studies of Organic Lake.

5.1.3 Organic lake biogeochemistry

The models of carbon, nitrogen and sulphur cycling in Organic Lake were constructed based on the presence and abundance of known marker genes along with data of the lake's chemical properties (chapter 4). As yet, metaproteomic analyses from the same samples have not been performed, but this would help corroborate the inferred pathways are active. Other inferences of Organic Lake biogeochemical function can be specifically tested on organisms in culture, tested by measuring rates of reaction *in situ* or supported with additional molecular and chemical analyses.

For the carbon cycle, it was hypothesized that aerobic anoxygenic photosynthesis (AAnP) and rhodopsin mediated photoheterotrophy can conserve carbon for use in biosynthesis reducing overall carbon loss (chapter 4). The conditions under which AAAnP is active can be tested on related organisms in culture, for example *Roseovarius tolerans* from Ekho Lake in the Vestfold Hills. It is already known for *R. tolerans* constant dim light suppresses bacteriochlorophyll A (BchlA) production while darkness stimulates it (Labrenz *et al.*, 1999). *R. tolerans*, or related isolates, can be grown under a larger range of conditions to determine when AAAnP becomes an active process. This might include further varying light intensity and cycle duration; varying organic carbon concentration and oxygen concentrations. To determine the contribution of light to growth, the difference in growth when AAAnP is active and when it is not can be compared. Characterising AAAnP in a model isolate from a similar Antarctic lake can inform an understanding of how it functions in Organic Lake. However, to gain an understanding of when AAAnP is active *in situ*, levels of BchlA at different points in the year can be measured to see it correlates with seasonal changes. In the case of the influence of rhodopsin-mediated phototrophy on the Organic Lake system, the function of rhodopsins first needs to be characterised in the diverse organisms in which they are found to confirm they are indeed used in photoheterotrophic growth. In marine *Flavobacteria*, there is already evidence that light stimulates growth through the action of rhodopsins (Gómez-Consarnau *et al.*, 2007). The same experimental design used to establish this can be applied to *Psychroflexus gondwanensis* isolated from Organic Lake, *Marinobacter* sp. ELB17 and *Octadecabacter antarcticus*, which are relatives of species found in Organic Lake.

In the model of the nitrogen cycle, nitrogen fixation was inferred to be negligible due to the high concentrations of ammonia while denitrification was assumed to be limited by low levels of oxidised nitrogen compounds and lack of potential for nitrification to regenerate them (chapter 4). In these cases where rates of reaction were inferred to be slow although genetic potential for them exists, the reaction rates can be directly

measured. Nitrogen fixation and denitrification rates can be measured from Organic Lake water samples by incubation with ^{15}N labelled substrates or acetylene block assays. Similarly, it was inferred that carbon fixation rates from photosynthetic algae and chemolithoautotrophic bacteria were lower than respiration and fermentation based on genetic potential (chapter 4). As mentioned previously, populations of chemolithoautotrophs may increase in the winter, potentially becoming significant source of primary production. Another consideration for the carbon budget is how sustained levels of chemolithoautotrophy throughout the year compares to the short burst of photosynthetic production over the summer. Continuous rates of chemolithoautotrophy appears to occur in Lake Bonney in McMurdo Dry Valleys as their expression of RuBisCO remained constant over the polar night transition (Kong *et al.*, 2012a). Measuring rates of primary production by ^{14}C incorporation and respiration rates at different points in the season can ascertain if there truly is a shortfall in the carbon budget in Organic Lake, estimate how large it is and pinpoint if the main source of primary production is from photo- or chemoautotrophy.

Both the carbon and nitrogen cycles were also constructed with the assumption that inputs are also negligible. For instance, denitrification rates would be limited by low fixation coupled with low rates of nitrification. Nonetheless, it is possible that sufficient nitrate inputs occur that would sustain denitrification thereby by-passing the need for endogenous nitrification. Monitoring for presence of melt streams that might feed into Organic Lake and determining their chemical profiles can gauge the contribution of allochthonous carbon and nitrogen to Organic Lake.

The unusually high concentrations of dimethylsulphide (DMS) in the bottom waters of Organic Lake were inferred to originate from bacterial lysis of dimethylsulphopropionate (DMSP) (chapter 4). As DMSP lyases have only been discovered fairly recently (Todd *et al.*, 2007; Curson *et al.*, 2008), they have only been characterised from a few isolates (Todd *et al.*, 2007; Curson *et al.*, 2008, 2010; Todd *et al.*, 2010; Curson *et al.*, 2012). Confirmation that the homologues found in Organic Lake are indeed functional can be achieved by cloning into the expression vector system established by Todd *et al.* (2007) and assaying for DMSP lyase activity. This would also provide valuable insight into the DMSP lyases that are most relevant to cold and hypersaline environments.

Although it was inferred that DMSP lysis was the main source of DMS in the bottom zone of Organic Lake, production of DMS other by anaerobic pathways are also possible. Furthermore, DMS was inferred to accumulate due to slow rates breakdown in the bottom zone. Incubating Organic Lake water samples with radio-labelled DMSP and tracking production of volatile DMS would confirm the lysis pathway is an active process and determine the rates of reaction. Performing this assay on water from different depths could determine if concentration of DMS is high in the bottom zone because it is being produced there at a higher rate. A similar assay can be performed using labelled DMS to show if DMS can be broken down and if this differs with depth. If the proposed model for sulphur cycling based on marker gene frequencies is correct, DMS breakdown in the bottom samples should be slow or not occur. While this experiment will not

exclude the possibility that DMS is produced by an alternative pathway, knowing the rates of DMSP lysis and DMS breakdown and the DMS concentration in a sample can indicate the existence other sources of DMS production.

Most of the Organic lake DMSP lyase genes could be linked to a taxonomic group except the most abundant type, OL-dddD, which had indication of both *Alpha-* or *Gammaproteobacteria* origins (chapter 4). Identifying the organisms involved in DMSP lysis can be achieved with SIP. This would involve incubation of ^{13}C -labelled DMSP in Organic Lake water samples followed by sequencing of the labelled DNA. Sequencing of the small subunit ribosomal RNA (SSU) genes would determine which specific taxonomic groups were the main contributors to DMSP lysis and screening for DMSP lyase genes would identify which DMSP lyase is involved. Performing this experiment at different depths of the water column could determine if the same organisms are responsible for DMSP lysis at different depths. No marker genes yet exist for DMS oxidation although it can be readily utilised as a growth substrate by diverse microorganisms (Johnston *et al.*, 2008). The same experiment can be performed using ^{13}C -labelled DMS to determine the taxa involved in breaking it down in different depths of the lake. In conjunction with metagenomic sequencing, or flow sorting and SAGs of DMS/DMSP-degraders, the biochemical pathways involved in DMS breakdown can then be reconstructed giving a more complete understanding of Organic Lake sulphur biogeochemistry.

5.2 Perspectives on ‘-omics’ approaches

It was not so long ago that what could be called the second molecular revolution in microbial ecology began in earnest with the first shotgun metagenomic sequencing of a marine virome (Breitbart *et al.*, 2002). That metagenome, sequenced with Sanger sequencing technology, was a modest 1.28 Mbp (Breitbart *et al.*, 2002). The first shotgun sequenced metagenomes of cellular life from the Sargasso Sea (Venter *et al.*, 2004) and the Iron Mountain acid mine drainage (Tyson *et al.*, 2004) added 265 Mbp and 76 Mbp respectively to the public databanks.

5.2.1 Next generation sequencing technologies

Since the availability of high-throughput next generation sequencing (NGS) technologies, the volume of metagenomic data has increased exponentially. To put it in perspective, a single sample from the Antarctic lake datasets described in this thesis sequenced by the Roche GS-FLX titanium sequencer is 140 Mbp, while a recent study using Applied Biosystems SOLiD sequencing of a marine sample (Iverson *et al.*, 2012) analysed 55,000 Mbp – close to 400 times the amount of data. This illustrates how sequencing technology is advancing so rapidly that each new study has to develop new ways of looking into the microbial milieu captured in the particular sample.

Current NGS sequencing platforms are shown in Table 5.1 comparing sequence length, throughput and error rates. Clearly, there are trade-offs between each type

Table 5.1: Comparison of next generation DNA sequencing platforms from Scholz *et al.* (2012). Indel, insertion-deletion; Sub, substitution; MP, mate pair; PE, paired end.

| Platform | Run time (h) | Read length (bp) | Mbp/run | Error type | Error rate (%) |
|----------------------------|------------------|--------------------------|---------------|------------|----------------|
| <i>Roche</i> | | | | | |
| 454 FLX+ | 18–20 | 700 | 900 | Indel | 1 |
| 454 FLX+ Titanium | 10 | 400 | 500 | Indel | 1 |
| 454 GS | 10 | 400 | 50 | Indel | 1 |
| <i>Illumina</i> | | | | | |
| GAIIX | 14 | 2 × 150 | 96,000 | Sub | >0.1 |
| HiSeq 2000 | 8 | 2 × 100 | 400,000 | Sub | >0.1 |
| HiSeq 2000 V3 | 10 | 2 × 150 | <600,000 | Sub | >0.1 |
| MiSeq | 1 | 2 × 150 | 1,000 | Sub | >0.1 |
| <i>Life Technologies</i> | | | | | |
| SOLiD 4 | 12 | 50 × 35 | 71,000 | AT bias | >0.06 |
| SOLiD 4 | 12 | 75 × 35 PE 60 × 60 MP | 155,000 | AT bias | >0.01 |
| <i>Ion Torrent</i> | | | | | |
| PGM 314 Chip | 3 | 100 | 10 | Indel | 1 |
| PGM 316 Chip | 3 | 100+ | 100 | Indel | 1 |
| PGM 318 Chip | 3 | 200 | 1,000 | Indel | 1 |
| <i>Pacific biosciences</i> | | | | | |
| RS | 14/8 Smart Cells | 1,500 | 45/smart cell | Insertions | 15 |

with higher read depth technologies offering shorter read lengths. Application of sequencing technologies that offer greater read depths in future Antarctic lake studies can enable insight into the rare members of the community. For example, a recent study compared SSU community profiles over a six year time course with a single sample deep-sequenced with the Illumina GAIIX platform (Caporaso *et al.*, 2012). Members of the community that appeared to become seasonally absent in the shallow sequenced community profiles were actually present in low abundances in the deep-sequenced sample (Caporaso *et al.*, 2012). The ability to look into deeper into diversity of the Antarctic lake communities can discern if a similar persistent ‘seed bank’ population exists in the lakes, if succession of populations occurs or external seeding. These three different possibilities have different implications for predicting responses to change in environmental conditions in the lakes.

5.2.2 Emerging bioinformatic bottlenecks

With the higher throughput and the falling costs of sequencing, genomic projects are now experiencing ‘bioinfomatic bottlenecks’ where computational analysis has become a significant challenge (reviewed by Scholz *et al.* (2012)). This is because availability of computational resources and scaling-up of algorithms to accomodate more data or computing clusters is not occurring as quickly as the accumulation of sequence data. Metagenomic sequencing projects typically entail computationally intensive BLAST-like comparisons before biological interpretations can be made. Local cluster computing, like that used in this thesis, has been a standard way to process metagenomic datasets. However, the annual cost of acquiring, running and administering a standard

rack mount server has been estimated to be \$2,160–\$5,160 US per node (Wilkening *et al.*, 2009). Acquiring and maintaining the necessary computational resources to process large datasets can in some cases work out to be more expensive than the cost of sequencing itself. Moreover, the time to acquire sequence data is much shorter than the time needed to analyse it. For example, analysis of metagenomic data produced by Illumina technology using a standard pipeline was calculated to take decades on a single processor or weeks to months on 1,000 CPUs (Evanko, 2009). Without increases in compute resources, the time taken to process massive NGS datasets will become prohibitively slow and is an important consideration for future metagenomic projects.

5.2.3 Prospects for closing the bioinformatic gap

Ultimately, for computational analysis to catch up with advances in sequencing technology requires more efficient processing, faster processors and/or faster algorithms. Ways to surpass the bioinformatic bottleneck are being developed and, if paired with careful experimental and analytical design, can make this a manageable challenge for prospective studies.

Webserver analyses

A shortage in computational resources can be addressed to some extent by the use of webserver pipelines specialised in metagenomic processing. These include Metagenome Rapid Annotation using Subsystems Technology (MG-RAST) (Meyer *et al.*, 2008), Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (Sun *et al.*, 2011), Integrated Microbial Genomes and Metagenomes (IMG/M) (Markowitz *et al.*, 2008, 2012), EBI Metagenomics (Hunter *et al.*, 2012), as well as METAVIR (Roux *et al.*, 2011) and Viral Informatics Resource for Metagenome Exploration (VIROME) (Wommack *et al.*, 2012) for viral metagenomes. These web-servers take on the burden of maintaining necessary programs, storage of sequence databases and provide computational power. Each have unique features that may be more desirable for particular sample types, analyses or user preferences. One disadvantage to web-servers is workflows are not readily customisable. The exception to this is WebGMA (Wu *et al.*, 2011), which hosts the most common metagenomic tools and enables them to be put together into custom pipelines, but does not contain a repository of metagenomic sequences. Current web-servers have been designed towards analysis of either bacteria/archaea or viral sequences; although the metagenomes from this thesis contained significant proportions of both, as well as eucaryotic sequences. To take advantage of the range of specialist features offered by different pipelines can introduce inefficiencies such as performing similar sequence quality control steps in order to fit within their distinct analytical schemas. The different outputs then need to be combined locally and may not be directly comparable to one another due to differences in processing. Nonetheless, use of webserver pipelines is a worthwhile option in future metagenomic studies of the lakes, especially if specialist tools or specific databases are needed, and can relieve some pressure on local computing time.

Cloud computing and GPUs

Cloud computing, the use of computing infrastructure, software or platforms as a service, looks to be a promising way to make up shortfalls in computational resources by affording large amounts of readily accessible compute time (reviewed by Schatz *et al.* (2010) and Thakur *et al.* (2012)). Several major providers exist: Amazon's Elastic Compute Cloud (EC2) is the largest publicly available service and the Department of Energy's Magellen Cloud is available for academic use. Gains in efficiency for the user can occur because clouds provide access to computing power that varies 'elastically' with the demands of the task without the cost of maintaining a local cluster. These infrastructure costs are more effectively absorbed by large providers than by smaller facilities enabling clouds to provide the same resources at lower cost. Also, by having access to more computational nodes, programs like BLAST can be run in parallel increasing the throughput of analysis.

Whether it works out to be more cost efficient in practice depends of the expected volume of data and frequency of analysis (Wilkening *et al.*, 2009). The cost–benefit appears greatest if use is sporadic or if computer resources need to be augmented on an unpredictable basis (Wilkening *et al.*, 2009). To benchmark the costs and time involved in metagenomic data analysis on the EC2, Angiuoli *et al.* (2011) clustered and BLAST compared 631 Mbp of metagenomic sequence (454 GS FLX titanium) against RefSeq and COG databases. This analysis finished in a little over 6 hours using a maximum of 160 CPUs at a total cost of \$56 US (Angiuoli *et al.*, 2011), giving an indication of how clouds might be leveraged for metagenomic processing. However, additional considerations such as adequate connection speeds to upload large quantities of data, data security, data management and computing expertise must also be taken into account. Cloud resources look to be most useful in future Antarctic projects for running computationally intensive analyses, such as BLAST searches and *de novo* assembly (discussed below).

Cutting down processing times of NGS data analysis can also be achieved by hardware-driven solutions. The most promising of these is use of graphical processing units (GPUs), which enable faster processing than a conventional central processing units (CPUs). A trial comparing the use of GPUs for processing Illumina GAIIx metagenomic datasets has reported 10–15 fold increase in processing speed compared to a CPU cluster (Su *et al.*, 2012). Currently, use of GPUs seems to be limited only by availability of bioinformatics software that can run on the different architecture and access to GPU clusters. A specific metagenomic analysis pipeline called Parallel-META has been launched that can run 16S rRNA prediction from a shotgun metagenome, classify the 16S sequences by BLAST and perform comparisons of multiple samples (Su *et al.*, 2012). More pipelines such as these and increasing availability of GPU clusters will undoubtedly be useful for speeding up metagenomic analyses.

Bioinformatic tools for binning and assembly of NGS data

The increased output of NGS sequencing technology is spurring development of bioinformatic programs which can solve specific metagenomic computational problems and decrease processing time. Metagenomic analyses are typically based on data from single reads, (*e.g.* chapter 4) or from assembled genomic information (*e.g.* chapter 3). These analyses entail two tasks that are unique to metagenomics: classification/binning of short sequences and assembly from mixed species datasets.

Broadly, programs for classifying reads use homology-based methods, where reads are compared to known sequences, or composition-based methods, where reads are clustered according to intrinsic properties. Homology-based programs, such as MEGAN (Huson *et al.*, 2007), CARMA (Krause *et al.*, 2008), WebCARMA (Gerlach *et al.*, 2009) and TREEPHYLER, are accurate and can be used with short reads but require computationally intensive BLAST or Hidden Markov Model searches (HMMer) against known sequence databases. The latest homology-based algorithms for short reads (<100 bp) have reduced processing time by using faster search algorithms. These are METABIN (Sharma *et al.*, 2012), which claims several-fold improvement in speed by implementing BLAT search (Kent, 2002), and GENOMETA (Davenport *et al.*, 2012), which reports an order of magnitude improvement in speed by using BOWTIE. As homology-based approaches all rely on a sequence databases, their ability to discover novel taxa is limited by the database. Compositional-based tools on the other hand allow reads with no sequenced homologues to be categorised, and they can generally be run on a laptop computer. Some examples of popular programs are TETRA (Teeling *et al.*, 2004) (tetra nucleotide-based), PHYLOPYTHIA (machine-learning based) (McHardy *et al.*, 2006) COMPOSTBIN (PCA-based) and TACOA (kernelized nearest-neighbor) (Diaz *et al.*, 2009). These older methods can only achieve high accuracy with longer reads (300–800 bp) due to local variations in genomic composition (reviewed in Teeling and Glöckner (2012)). The latest binning algorithms, ABUNDANCEBIN that uses l-tuples (Wu and Ye, 2011) and the mixture models developed by Meinicke *et al.* (2011), are a lot less sensitive to read length and can be used effectively for binning of reads produced by Illumina platforms. Programs have also been developed that take a combined homology- and composition-based approach including PhymmBL (Brady and Salzberg, 2009; Rosen *et al.*, 2011) and the NB-based classifier developed by Parks *et al.* (2011).

The deep sequencing capability of NGS technologies enables increased coverage of genomes in the environment and therefore a greater possibility for complete genomes to be assembled *de novo*. However, *de novo* genome assembly from a high volume of short reads is an extremely difficult computational problem requiring a great deal of computer memory to run (see Salzberg *et al.* (2012) for a review of assembly algorithms). Application of assembly programs designed for single genomes produced chimeras when applied to simulated metagenomic sequences questioning their applicability to mixed species data (Pignatelli and Moya, 2011). Several *de novo* assemblers have been released specifically for metagenomic sequences, namely META-IBDA (Peng *et al.*, 2011), GENOVO (Laserson *et al.*, 2011) and METAVELVET (Namiki *et al.*, 2012).

Given the large memory requirements of these programs, it remains to be seen how they will perform in large sequencing projects.

Two recent developments hold promise for improved *de novo* assembly of metagenomes sequenced to high read depth. One study has assembled a near-complete *Rhodobacterales* genome and two variants of marine group II *Euryarchaeota* genomes from SOLiD sequenced marine metagenomes (Iverson *et al.*, 2012). Surprisingly, these taxa only comprised <10% of the community. Assembly of metagenomic data using VELVET only produced short contigs. Extensive post-processing using in-house software, SEASTAR and TETRACALC, were necessary to further assemble these contigs into scaffolds. This involved forming scaffolds based on coverage, GC content and matepairing scores and binning those scaffolds into genomes based on nucleotide composition statistics. Software used to generate these genomes will be made available in the future at <http://armbrustlab.ocean.washington.edu/seastar>, which will allow the assembly methodology to be validated on other metagenomic samples and assessed for broad applicability. In particular, information about the computational resources and degree of manual curation necessary to achieve this sort of assembly is needed to gauge its utility. The other recent study has developed a method for overcoming memory limitations when using NGS data assemblers, such as VELVET, on metagenomic data (Pell *et al.*, 2012). In this approach, the *k*-mer connectivity (de Bruijn) graphs produced by the assembler are stored as a simplified representation based on Bloom filters requiring less memory. The representations of the de Bruijn graphs were used to partition reads into smaller connected sets. When applied to a Illumina GII sequenced soil metagenome, assembly was identical when using partitioned and unpartitioned data but with a 40-fold decrease in memory usage in the partitioned dataset. Again, further testing is required to evaluate broad applicability of this method but it looks to be scalable for any size dataset.

Future Antarctic lake metagenomics projects can profit immensely from the wealth of emerging bioinformatics tools available. Complete viral genomes were able to be assembled *de novo* from the 2006 Organic Lake samples using a combination of Sanger and 454 data (chapter 3) demonstrating assembly-driven approaches are effective in Antarctic lakes. This suggests future studies utilising deep sequencing and new assembly algorithms could obtain complete genome sequences and from less abundant members of the community. However, no single bioinformatic pipeline will cover all aspects of metagenomic analysis and as sequencing technologies are advancing rapidly, there is no established best practice at the moment for metagenomic analyses. Deciding on which analysis approach to take will require careful assessment of the sequence data and bioinformatic resources available.

5.2.4 'omes are only as good as our databases

Aside from the computational limitations, '-omics' type analyses rely a great deal on current sequence databases and laboratory studies to make meaningful inferences. For example, detection of the OLV (chapter 3) was contingent on the presence of a suffi-

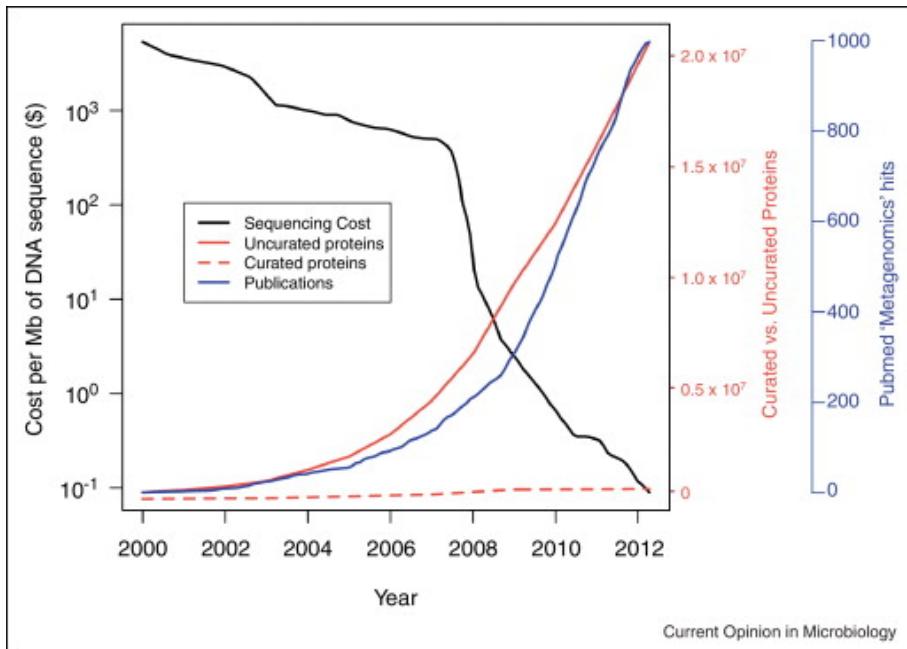


Figure 5.1: Plot comparing sequencing cost, uncurated protein data, curated protein data and metagenomics publications from 2000 to 2012 from Temperton and Giovannoni (2012). Sequencing costs retrieved from <http://www.genome.gov/sequencingcosts/>. Uncurated proteins were the total number of sequences in the UniProt TrEMBL database. Curated proteins were the total the manually annotated entries in the SwissProt database. Metagenomic publications were the total retrieved from PubMed using the search term ‘metagenomics’.

ciently close relative, Sputnik (La Scola *et al.*, 2008), in the non-redundant database. Without the Sputnik sequence in the database, the OLV scaffold would still have been assembled, but it would have had no capsid protein match to identify the scaffold as viral. Furthermore, without the physiological characterisation of Sputnik as a virophage of mimivirus (La Scola *et al.*, 2008), the genomic inferences linking OLV to OLPV would not have been possible demonstrating how a biological knowledge-base is also crucial to understand metagenomic data. Similarly, it was due to the recent sequencing and characterisation of DMSP lyase genes that they were specifically sought out in the Organic Lake metagenomic sequences as a marker for DMS production (chapter 4).

A dependence on sequence databases applies even more to metaproteomic analysis. Unlike metagenomics where fairly distantly related sequences can be matched to a metagenome to give some idea of its biological significance, spectral matching fails if protein sequences with high identity to those in the sample are not available (chapter 2). This makes metaproteomics much more reliant on sequence database availability and accuracy. Furthermore, identification of proteins only indicates that they are expressed, but characterisation of the same or closely related protein is still necessary to infer its function in the sample. Knowledge of protein function is accumulating extremely slowly compared to sequence data (Figure 5.1). This illustrates how both metagenomics and metaproteomics need to be paired with basic laboratory studies for meaningful biological inferences continue to be made. Conversely, it also means a great

deal of metagenomics sequences and metaproteomic mass spectra data remains to be taxonomically or functionally assigned.

The specific hypotheses built on environmental data can be used to guide laboratory research studies, such as were detailed above in *Possible future work on Organic Lake*, thus feeding back into supporting a systems-level understanding of microbial communities. Future ‘-omic’ studies need to now encompass longer time scales and employ focussed analysis of populations and individuals to build on known sequence databases and our biological knowledge base. Combining ‘-omics’ tools with laboratory work on isolates and with complementary techniques for targeting specific populations, community functions and biogeochemical processes holds promise for enabling a deeper understanding of ecosystems.

5.2.5 Metagenomes and metaproteomes are time capsules

The DNA sequence and mass spectra data that has been produced form a lasting record of the state of the microbial communities at the point in time that they were collected. It thus acts as a resource against which other datasets can be compared to gain an understanding of variability between ecosystems and as a bench mark to gauge changes overtime. The data can also be specifically mined to recover features of interest, for example sequences that encode enzymes with desired activities or bioactive compounds can be recovered by synthesis from the sequence information. As sequence databases grow and more of the microbial world is characterised, this archive of molecular data can be re-analysed to learn even more from these unique ecosystems.

5.3 Concluding remarks

This thesis has demonstrated how metagenomics and metaproteomics can be used as a tool for characterising Antarctic lake ecosystems.

References

- Abell G. C. J. and Bowman J. P. (2005). Ecological and biogeographic relationships of class *Flavobacteria* in the Southern Ocean. *FEMS Microbiology Ecology*, 51: 265–77.
- Abell G. C. J. and Bowman J. P. (2005). Colonization and community dynamics of class *Flavobacteria* on diatom detritus in experimental mesocosms based on Southern Ocean seawater. *FEMS Microbiology Ecology*, 53:379–391.
- Agarkova I. V., Dunigan D. D., and Van Etten J. L. (2006). Virion-associated restriction endonucleases of chloroviruses. *Journal of Virology*, 80:8114–8123.
- Alekhina I. A., Marie D., Petit J. R., Lukin V. V., Zubkov V. M., and Bulat S. A. (2007). Molecular analysis of bacterial diversity in kerosene-based drilling fluid from the deep ice borehole at Vostok, East Antarctica. *FEMS Microbiology Ecology*, 59: 289–299.
- Allen L. Z., Ishoey T., Novotny M. A., McLean J. S., Lasken R. S., and Williamson S. J. (2011). Single virus genomics: a new tool for virus discovery. *PLoS One*, 6: e17722.
- Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.
- Angiuoli S. V., White J. R., Matalka M., White O., and Fricke W. F. (2011). Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS One*, 6:e26624.
- Balashov S. P., Imasheva E. S., Boichenko V. A., Antón J., Wang J. M., and Lanyi J. K. (2005). Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. *Science*, 309:2061–2064.
- Béjà O., Suzuki M. T., Heidelberg J. F., Nelson W. C., Preston C. M., Hamada T., Eisen J. A., Fraser C. M., and Delong E. F. (2002). Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature*, 415:5–8.
- Bell E. M. and Laybourn-Parry J. (2003). Mixotrophy in the Antarctic phytoflagellate, *Pyramimonas gelidicola* (*Chlorophyta*: *Prasinophyceae*). *Journal of Phycology*, 649:644–649.
- Bengtsson J., Eriksson K. M., Hartmann M., Wang Z., Shenoy B. D., Grelet G.-A., Abarenkov K., Petri A., Rosenblad M. A., and Nilsson R. H. (2011). Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*, 100:471–475.

- Bergh O., Borsheim K. Y., Bratbak G., and Heldal M. (1989). High abundance of viruses found in aquatic environments. *Nature*, 340:467–468.
- Bergmann D. J., Hooper A. B., and Klotz M. G. (2005). Structure and sequence conservation of *ji_ihao_j/i_k* cluster genes of autotrophic ammonia-oxidizing bacteria: evidence for their evolutionary history. *Applied and Environmental Microbiology*, 71: 5371–5382.
- Bielewicz S., Bell E. M., Kong W., Friedberg I., Priscu J. C., and Morgan-Kiss R. M. (2011). Protist diversity in a permanently ice-covered Antarctic lake during the polar night transition. *The ISME Journal*, 5:1559–1564.
- Bird M. I., Chivas A. R., Radnell C. J., and Burton H. R. (1991). Sedimentological and stable-isotope evolution of lakes in the Vestfold Hills, Antarctica. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 84:109–130.
- Bouvier T. and Giorgio P.del (2007). Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environmental Microbiology*, 9:287–297.
- Bowman J. P., Mccammon S. A., Lewis T., Skerratt J. H., Brown J. L., Nichols D. S., and Mcmeekin T. A. (1998). *ji_iPsychroflexus torquis_j/i_k* gen. nov., sp. nov., a psychrophilic species from Antarctic sea ice, and reclassification of *ji_iFlavobacterium gondwanense_j/i_k* (Dobson et al. 1993) as *ji_iPsychroflexus gondwanense_j/i_k* gen. nov., comb. nov. *Microbiology*, 144:1601–1609.
- Bowman J. P., McCammon S. A., Rea S. M., and McMeekin T. A. (2000). The microbial composition of three limnologically disparate hypersaline Antarctic lakes. *FEMS Microbiology Letters*, 183:81–88.
- Bowman J. P., Rea S. M., McCammon S. A., and McMeekin T. A. (2000). Diversity and community structure within anoxic sediment from marine salinity meromictic lakes and a coastal meromictic marine basin, Vestfold Hills, Eastern Antarctica. *Environmental Microbiology*, 2(2):227–237.
- Boyer M., Azza S., Barrassi L., Klose T., Campocasso A., Pagnier I., Fournous G., Borg A., Robert C., Zhang X., Desnues C., Henrissat B., Rossmann M. G., La Scola B., and Raoult D. (2011). Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proceedings of the National Academy of Sciences USA*, 108: 10296–102301.
- Brady A. and Salzberg S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6:673–676.
- Bratbak G., Jacobsen A., and Heldal M. (1998). Viral lysis of *ji_iPhaeocystis pouchetii_j/i_k* and bacterial secondary production. *Aquatic Microbial Ecology*, 16: 11–16.
- Breitbart M., Salamon P., Andresen B., Mahaffy J. M., Segall A. M., Mead D., Azam F., and Rohwer F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences USA*, 99:14250–14255.
- Bronge C. (2004). Hydrographic and climatic changes influencing the proglacial Druzhby drainage system, Vestfold Hills, Antarctica. *Antarctic Science*, 8:379–388.

- Budinoff C. R., Loar S. N., LeCleir G. R., Wilhelm S. W., and Buchan A. (2011). A protocol for enumeration of aquatic viruses by epifluorescence microscopy using Anodisc 13 membranes. *BMC Microbiology*, 11:168.
- Bulat S. A., Alekhina I. A., Blot M., Petit J.-R., Angelis M.de, Wagenbach D., Lipenkov V. Y., Vasilyeva L. P., Wloch D. M., Raynaud D., and Lukin V. V. (2004). DNA signature of thermophilic bacteria from the aged accretion ice of Lake Vostok, Antarctica: implications for searching for life in extreme icy environments. *International Journal of Astrobiology*, 3:1–12.
- Burke C. and Burton H. R. (1988). Photosynthetic bacteria in meromictic lakes and stratified fjords of the Vestfold Hills, Antarctica. *Hydrobiologia*, 165:13–23.
- Burton H. R. (1981). Chemistry, physics and evolution of Antarctic saline lakes. *Hydrobiologia*, 82:339–362.
- Campbell B. J., Engel A. S., Porter M. L., and Takai K. (2006). The versatile γ -Epsilon-proteobacteria: key players in sulphidic habitats. *Nature Reviews Microbiology*, 4:458–468.
- Caporaso J. G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F. D., Costello E. K., Fierer N., Peña A. G., Goodrich J. K., Gordon J. I., Huttley G. A., Kelley S. T., Knights D., Koenig J. E., Ley R. E., Lozupone C. A., McDonald D., Muegge B. D., Pirrung M., Reeder J., Sevinsky J. R., Turnbaugh P. J., Walters W. A., Widmann J., Yatsunenko T., Zaneveld J., and Knight R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335–336.
- Caporaso J. G., Paszkiewicz K., Field D., Knight R., and Gilbert J. A. (2012). The Western English Channel contains a persistent microbial seed bank. *The ISME journal*, 6:1089–1093.
- Carver T. J., Rutherford K. M., Berriman M., Rajandream M.-A., Barrell B. G., and Parkhill J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics*, 21:3422–3423.
- Castberg T., Larsen A., Sandaa R.-A., Brussaard C. P., Egge J., Heldal M., Thyrhaug R., Hannen E.van, and Bratbak G. (2001). Microbial population dynamics and diversity during a bloom of the marine cocolithophorid Emiliania huxleyi . *Marine Ecology Progress Series*, 221:39–46.
- Charlson R. J., Lovelock J. E., Andreae M. O., and Warren S. G. (1987). Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature*, 326:655–661.
- Chen F., Mackey A. J., Stoeckert C. J., and Roos D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34:D363–D368.
- Chouari R., Le Paslier D., Daegelen P., Ginestet P., Weissenbach J., and Sghir A. (2005). Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. *Environmental Microbiology*, 7:1104–1115.
- Christner B. C., Mosley-Thompson E., Thompson L. G., and Reeve J. N. (2001). Isolation of bacteria and 16S rDNAs from Lake Vostok accretion ice. *Environmental Microbiology*, 3:570–577.

- Clarke K. and Gorley R. PRIMER V6: User Manual/Tutorial, 2006.
- Claverie J.-M. and Abergel C. (2009). Mimivirus and its virophage. *Annual review of genetics*, 43:49–66.
- Collins R. E., Rocap G., and Deming J. W. (2010). Persistence of bacterial and archaeal communities in sea ice through an Arctic winter. *Environmental Microbiology*, 12: 1828–1841.
- Cottrell M. T. and Kirchman D. L. (2009). Photoheterotrophic microbes in the Arctic Ocean in summer and winter. *Applied and Environmental Microbiology*, 75:4958–4966.
- Curran M. A., Jones G. B., and Burton H. R. (1998). Spatial distribution of dimethylsulphide and dimethylsulfoniopropionate in the Australasian sector of the Southern Ocean. *Journal of Geophysical Research*, 103:16677–16689.
- Curson A. R., Rogers R., Todd J. D., Brearley C. A., and Johnston A. W. (2008). Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in several marine alpha-proteobacteria and *Rhodobacter sphaeroides*. *Environmental Microbiology*, 10:757–767.
- Curson A. R., Sullivan M. J., Todd J. D., and Johnston A. W. (2010). Identification of genes for dimethyl sulfide production in bacteria in the gut of Atlantic Herring (*Clupea harengus*). *The ISME Journal*, 4:144–146.
- Curson A. R., Sullivan M. J., Todd J. D., and Johnston A. W. (2011). DddY, a periplasmic dimethylsulfoniopropionate lyase found in taxonomically diverse species of Proteobacteria. *The ISME Journal*, 5:1191–200.
- Curson A. R., Todd J. D., Sullivan M. J., and Johnston A. W. (2011). Catabolism of dimethylsulphoniopropionate: microorganisms, enzymes and genes. *Nature Reviews Microbiology*, 9:849–859.
- Curson A. R., Fowler E. K., Dickens S., Johnston A. W., and Todd J. D. (2012). Multiple DMSP lyases in the gamma-proteobacterium *Oceanimonas doudoroffii*. *Biogeochemistry*, 110:109–119.
- Danovaro R., Corinaldesi C., Dell'Anno A., Fuhrman J. A., Middelburg J. J., Noble R. T., and Suttle C. A. (2011). Marine viruses and global climate change. *FEMS Microbiology Reviews*, 35:993–1034.
- Davenport C. F., Neugebauer J., Beckmann N., Friedrich B., Kameri B., Kokott S., Paetow M., Siekmann B., Wieding-Drewes M., Wienhöfer M., Wolf S., Tümmler B., Ahlers V., and Sprengel F. (2012). Genometa - a fast and accurate classifier for short metagenomic shotgun reads. *PloS One*, 7:e41224.
- Davidson A. T. and Marchant H. J. (1992). Protist abundance and carbon concentration during a *Phaeocystis*-dominated bloom at an Antarctic coastal site. *Polar Biology*, 12:387–395.
- DeMaere M. Z., Lauro F. M., Thomas T., Yau S., and Cavicchioli R. (2011). Simple high-throughput annotation pipeline (SHAP). *Bioinformatics*, 27:2431–2432.
- Demergasso C., Escudero L., Casamayor E. O., Chong G., Balagué V., and Pedrós-Alió C. (2008). Novelty and spatio-temporal heterogeneity in the bacterial diversity of hypersaline Lake Tebenquiche (Salar de Atacama). *Extremophiles*, 12:491–504.

- Denef V. J., Shah M. B., Verberkmoes N. C., Hettich R. L., and Banfield J. F. (2007). Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *Journal of Proteome Research*, 6:3152–3161.
- Deprez P. P., Franzmann P. D., and Burton H. R. (1986). Determination of reduced sulfur gases in Antarctic lakes and seawater by gas chromatography after solid adsorbent preconcentration. *Journal of Chromatography*, 362:9–21.
- Desnues C., La Scola B., Yutin N., Fournous G., Robert C., Azza S., Jardot P., Monteil S., Campocasso A., Koonin E. V., and Raoult D. (2012). Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proceedings of the National Academy of Sciences USA*, 109:18078–18083.
- Diaz N. N., Krause L., Goesmann A., Niehaus K., and Nattkemper T. W. (2009). TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, 10:56.
- Diemer G. S., Kyle J. E., and Stedman K. M. Counting viruses using polycarbonate Track Etch membrane filters as an alternative to Anodisc membrane filters, 2012. URL http://www.web.pdx.edu/~kstedman/PCTE_virus_counting_protocol.pdf.
- Dobson S., James S., Franzmann P. D., and Mcmeekin T. A. (1991). A numerical taxonomic study of some pigmented bacteria isolated from Organic Lake, an antarctic hypersaline lake. *Archives of Microbiology*, 156:56–61.
- Eberlein K., Leal M., Hammer K., and Hickel W. (1985). Dissolved organic substances during a *Phaeocystis pouchetii* bloom in the German Bight (North Sea). *Marine Biology*, 89:311–316.
- Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32:1792–1797.
- Edwards K., Rogers D., Wirsen C., and McCollom T. (2003). Isolation and characterization of novel psychrophilic, neutrophilic, Fe-oxidizing, chemolithoautotrophic alpha- and gamma-*Proteobacteria* from the deep sea. *Applied and Environmental Microbiology*, 69:2906–2913.
- Elshahed M. S., Najar F. Z., Aycock M., Qu C., Roe B. A., and Krumholz L. R. (2005). Metagenomic analysis of the microbial community at Zodletone Spring (Oklahoma): insights into the genome of a member of the novel candidate division OD1. *Applied and Environmental Microbiology*, 71:7598–7602.
- Evanko D. (2009). Metagenomics versus Moores law. *Nature Methods*, 6:623.
- Ferris J. M., Gibson J. A., and Burton H. R. (1991). Evidence of density currents with the potential to promote meromixis in ice-covered saline lakes. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 84:99–107.
- Fischer M. G. (2011). Sputnik and Mavirus: more than just satellite viruses. *Nature Reviews*, 10:1–2.
- Fischer M. G. and Suttle C. A. (2011). A virophage at the origin of large DNA transposons. *Science*, 332:231–234.

- Fischer M. G., Allen M. J., Wilson W. H., and Suttle C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences USA*, 107:19508–19513.
- Florens L., Carozza M. J., Swanson S. K., Fournier M., Coleman M. K., Workman J. L., and Washburn M. P. (2006). Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods*, 40:303–311.
- Fofonoff N. and Millard R. J. (1983). Algorithms for computation of fundamental properties of seawater. *UNESCO Technical Papers in Marine Science*, 44.
- Franzmann P. D., Burton H. R., and Mcmeekin T. A. (1987). *Halomonas subglaciescola*, a new species of halotolerant bacteria isolated from Antarctica. *International Journal of Systematic Bacteriology*, 37:27–34.
- Franzmann P. D., Deprez P. P., Burton H. R., and Hoff J.van den (1987). Limnology of Organic Lake, Antarctica, a meromictic lake that contains high concentrations of dimethyl sulfide. *Marine and Freshwater Research*, 38:409–417.
- Franzmann P. D., Stackebrandt E., Sanderson K., Volkman J. K., Cameron D., Stevenson P., and Mcmeekin T. A. (1988). *Halorubrum lacusprofundi*, sp. nov., a halophilic bacterium isolated from Deep Lake, Antarctica. *Systematic & Applied Microbiology*, 11:20–27.
- Friedrich C. G., Bardischewsky F., Rother D., Quentmeier A., and Fischer J. (2005). Prokaryotic sulfur oxidation. *Current Opinion in Microbiology*, 8:253–259.
- Fuchs B. M., Spring S., Teeling H., Quast C., Wulf J., Schattenhofer M., Yan S., Ferriera S., Johnson J., Glöckner F. O., and Amann R. (2007). Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. *Proceedings of the National Academy of Sciences USA*, 104:2891–2896.
- Fuhrman J. A., Schwalbach M. S., and Stingl U. (2008). Proteorhodopsins: an array of physiological roles? *Nature Reviews Microbiology*, 6:488–494.
- Gärdes A., Kaeppele E., Shehzad A., Seebah S., Teeling H., Yarza P., Glöckner F. O., Grossart H.-P., and Ullrich M. S. (2010). Complete genome sequence of *Marinobacter adhaerens* type strain (HP15), a diatom-interacting marine microorganism. *Standards in Genomic Sciences*, 3:97–107.
- Gauthier M., Lafay B., Christen R., Fernandez L., Acquaviva M., Bonin P., and Bertrand J.-C. (1992). A new, extremely halotolerant, hydrocarbon-degrading marine bacterium. *International Journal of Systematic Bacteriology*, pages 568–576.
- Gerlach W., Jünemann S., Tille F., Goesmann A., and Stoye J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10:430.
- Gibson J. A. (1999). The meromictic lakes and stratified marine basins of the Vestfold Hills, East Antarctica. *Antarctic Science*, 11:175–192.
- Gibson J. A. and Burton H. R. (1996). Meromictic Antarctic lakes as recorders of climate change: the structures of Ace and Organic Lakes, Vestfold Hills, Antarctica. *Papers and Proceedings of the Royal Society of Tasmania*, 130:73–78.

- Gibson J. A., Ferris J. M., and Burton H. R. (1990). Temperature density, temperature conductivity and conductivity-density relationships for marine-derived saline lake waters. *ANARE Research Notes*, 78.
- Gibson J. A., Garrick R. C., Franzmann P. D., Deprez P. P., and Burton H. R. (1991). Reduced sulfur gases in saline lakes of the Vestfold Hills, Antarctica. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 84:131–140.
- Gibson J. A., Qiang X. L., Franzmann P. D., Garrick R. C., and Burton H. R. (1994). Volatile fatty and dissolved free amino acids in Organic Lake, Vestfold Hilss, East Antarctica. *Polar Biology*, 14:545–550.
- Gibson J. A., Burton H. R., and Gallagher J. (1995). Meromictic Antarctic lakes as indicators of local water balance: structural changes in Organic Lake, Vestfold Hills 1978–1994. *ANARE Research Notes*, 94:16–18.
- Glatz R., Lepp P., Ward B. B., and Francis C. (2006). Planktonic microbial community composition across steep physical/chemical gradients in permanently ice-covered Lake Bonney, Antarctica. *Geobiology*, 4:53–67.
- Goberna M., Insam H., and Franke-Whittle I. (2009). Effect of biowaste sludge maturation on the diversity of thermophilic bacteria and archaea in an anaerobic reactor. *Applied and Environmental Microbiology*, 75:2566–2572.
- Gómez-Consarnau L., González J. M., Coll-Lladó M., Gourdon P., Pascher T., Neutze R., Pedrós-Alio C., and Pinhassi J. (2007). Light stimulates growth of proteorhodopsin-containing marine *Flavobacteriia*. *Nature*, 445:210–213.
- Gómez-Consarnau L., Akram N., Lindell K., Pedersen A., Neutze R., Milton D. L., González J. M., and Pinhassi J. (2010). Proteorhodopsin phototrophy promotes survival of marine bacteria during starvation. *PLoS Biology*, 8:2–11.
- Gordon D., Priscu J. C., and Giovanonni S. J. (2000). Origin and phylogeny of microbes living in permanent Antarctic lake ice. *Microbial Ecology*, 39:197–202.
- Gordon D. Viewing and editing assembled sequences using Consed. In Baxvanis A. and Davison D., editors, *Current Protocols in Bioinformatics*, volume Chapter 11, chapter Viewing an, pages 11.2.1–11.2.43. John Wiley & Sons, New York, 2004. doi: 10.1002/0471250953.bi1102s02. URL <http://www.ncbi.nlm.nih.gov/pubmed/19664765>.
- Green D. H., Bowman J. P., Smith E. A., Gutierrez T., and Bolch C. J. (2006). *Marinobacter algicola* sp. nov., isolated from laboratory cultures of paralytic shellfish toxin-producing dinoflagellates. *International Journal of Systematic and Evolutionary Microbiology*, 56:523–527.
- Green W. J., Angle M. P., and Chave K. E. (1988). The geochemistry of Antarctic streams and their role in the evolution of four lakes of the McMurdo Dry Valleys. *Geochimica et Cosmochimica Acta*, 52:1265–1274.
- Grzymski J. J., Riesenfeld C. S., Williams T. J., Dussaq A. M., Ducklow H., Erickson M., Cavicchioli R., and Murray A. E. (2012). A metagenomic assessment of winter and summer bacterioplankton from Antarctica Peninsula coastal surface waters. *The ISME Journal*, 6:1901–1915.
- Haft D. H., Loftus B., Richardson D., Yang F., Eisen J. A., Paulsen I. T., and White O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, 29:41–43.

- Hahn M. W. (2009). Description of seven candidate species affiliated with the phylum *Actinobacteria*, representing planktonic freshwater bacteria. *International Journal of Systematic and Evolutionary Microbiology*, 59:112–117.
- Hahn M. W., Lünsdorf H., Wu Q., Schauer M., Höfle M. G., Boenigk J., and Stadler P. (2003). Isolation of novel ultramicrobacteria classified as *Actinobacteria* from five freshwater habitats in Europe and Asia. *Applied and Environmental Microbiology*, 69:1442–1451.
- Hahn M. W., Stadler P., Wu Q. L., and Pöckl M. (2004). The filtration-acclimatization method for isolation of an important fraction of the not readily cultivable bacteria. *Journal of Microbiological Methods*, 57:379–390.
- Hahsler M., Hornik K., and Buchta C. (2007). Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software*, 25:1–34.
- Handelsman J. (2008). Metagenomics is not enough. *DNA and Cell Biology*, 27:219–221.
- Hara S., Terauchi K., and Koike I. (1991). Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. *Applied and Environmental Microbiology*, 57:2731–2734.
- Harris J. K., Kelley S. T., and Pace N. R. (2004). New perspective on uncultured bacterial phylogenetic division OP11. *Applied and Environmental Microbiology*, 70:845–849.
- Hennes K. P. and Suttle C. A. (1995). Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. *Limnology and Oceanography*, 40:1050–1055.
- Hobbie J., Daley R., and Jasper S. (1977). Use of nucleopore filters for counting bacteria by fluorescence microscopy. *Applied and Environmental Microbiology*, 33:1225–1228.
- Hodgson D. A. Antarctic Lakes. In Bengtsson L., Herschy R. W., and Fairbridge R. W., editors, *Encyclopedia of Lakes and Reservoirs*, pages 26–31. Springer, Dordrecht, 2012.
- Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N., Singhal M., Xu L., Mendes P., and Kummer U. (2006). COPASI - a COmplex PAthway SImlator. *Bioinformatics*, 22:3067–3074.
- Horvath P. and Barrangou R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science*, 327:167–170.
- Howard E. C., Henriksen J. R., Buchan A., Reisch C. R., Bürgmann H., Welsh R., Ye W., González J. M., Mace K., Joye S. B., Kiene R. P., Whitman W. B., and Moran M. A. (2006). Bacterial taxa that limit sulfur flux from the ocean. *Science*, 314:649–652.
- Howard E. C., Sun S., Biers E. J., and Moran M. A. (2008). Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environmental Microbiology*, 10:2397–2410.
- Huang L.-N., Zhu S., Zhou H., and Qu L.-H. (2005). Molecular phylogenetic diversity of bacteria associated with the leachate of a closed municipal solid waste landfill. *FEMS Microbiology Letters*, 242:297–303.

- Huang Y., Niu B., Gao Y., Fu L., and Li W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26:260–262.
- Hügler M. and Sievert S. M. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Annual Review of Marine Science*, 3:261–289.
- Humayoun S. B., Bano N., and Hollibaugh J. T. (2003). Depth distribution of microbial diversity in Mono Lake, a meromictic soda lake in California. *Applied and Environmental Microbiology*, 69:1030–1042.
- Hunter C. I., Mitchell A., Jones P., McAnulla C., Pesseat S., Scheremetjew M., and Hunter S. (2012). Metagenomic analysis: the challenge of the data bonanza. *Briefings in Bioinformatics*, 13:743–746.
- Huson D. H., Auch A. F., Qi J., and Schuster S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17:377–386.
- Huu N. B., Denner E. B., Ha D. T., Wanner G., and Stan-Lotter H. (1996). *Marinobacter aquaeolei* sp. nov., a halophilic bacterium isolated from a Vietnamese oil-producing well. *International Journal of Systematic Bacteriology*, 49: 367–375.
- Iverson V., Morris R. M., Frazer C. D., Berthiaume C. T., Morales R. L., and Armbrust E. V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine *Euryarchaeota*. *Science*, 335:587–590.
- Iyer L. M., Aravind L., and Koonin E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *Journal of Virology*, 75:11720–11734.
- Iyer L. M., Balaji S., Koonin E. V., and Aravind L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research*, 117:156–184.
- Jacobsen A., Bratbak G., and Heldal M. (1996). Isolation and characterization of a new virus infecting *Phaeocystis pouchetii* (Prymnesiophyte). *Journal of Phycology*, 32:586–588.
- James S., Dobson S., Franzmann P. D., and Mcmeekin T. A. (1990). *Halomonas meridianai*, a new species of extremely halotolerant bacteria isolated from Antarctic saline lakes. *Systematic and Applied Microbiology*, 13(3):270–278.
- James S., Burton H. R., Mcmeekin T. A., and Mancuso C. (1994). Seasonal abundance of *Halomonas meridianai*, *Halomonas subglaciescens*, *Flavobacterium gondwanense* and *Flavobacterium salegens* in four Antarctic lakes. *Antarctic Science*, 6:325–332.
- Johnston A. W., Todd J. D., Sun L., Nikolaidou-Katsaraidou N., Curson A. R., and Rogers R. (2008). Molecular diversity of bacterial production of the climate-changing gas, dimethyl sulphide, a molecule that impinges on local and global symbioses. *Journal of Experimental Botany*, 59:1059–1067.
- Johnstone G., Brown D., and Lugg D. (1973). The biology of the Vestfold Hills, Antarctica. *ANARE Scientific Reports*, 123:1–60.
- Kang I., Lee K., Yang S.-J., Choi A., Kang D., Lee Y. K., and Cho J.-C. (2012). Genome sequence of "Candidatus Aquiluna" sp. strain IMCC13023, a marine member of the *Actinobacteria* isolated from an Arctic fjord. *Journal of Bacteriology*, 194:3550–3551.

- Karginov F. V. and Hannon G. J. (2010). The CRISPR system: small RNA-guided defense in bacteria and archaea. *Molecular Cell*, 37:7–19.
- Karr E. A., Sattley W. M., Jung D. O., Madigan M. T., and Achenbach L. A. (2003). Remarkable diversity of phototrophic purple bacteria in a permanently frozen Antarctic lake. *Applied and Environmental Microbiology*, 8:4910–4914.
- Karr E. A., Sattley W. M., Rice M. R., Jung D. O., Madigan M. T., and Achenbach L. A. (2005). Diversity and distribution of sulfate-reducing bacteria in permanently frozen Lake Fryxell, McMurdo Dry Valleys, Antarctica. *Applied and Environmental Microbiology*, 71:6353–6359.
- Karr E. A., Ng J. M., Belchik S. M., Matthew W. M., Madigan M. T., Achenbach L. A., and Sattley W. M. (2006). Biodiversity of methanogenic and other *ii₃Archaea*ii** in the permanently frozen Lake Fryxell, Antarctica. *Applied and Environmental Microbiology*, 72(2):1663–1666.
- Keller A., Nesvizhskii A. I., Kolker E., and Aebersold R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74:5383–5392.
- Kent W. J. (2002). BLATthe BLAST-like alignment tool. *Genome Research*, 12:656–664.
- Kirchman D. L. (2002). The ecology of *ii₃Cytophaga-Flavobacteriia*ii** in aquatic environments. *FEMS Microbiology Ecology*, 39:91–100.
- Kong W., Dolhi J. M., Chiuchiolo A., Priscu J., and Morgan-Kiss R. M. (2012). Evidence of form II RubisCO (cbbM) in a perennially ice-covered Antarctic lake. *FEMS Microbiology Ecology*, 82:1–10.
- Kong W., Ream D. C., Priscu J. C., and Morgan-Kiss R. M. (2012). Diversity and expression of RubisCO genes in a perennially ice-covered Antarctic lake during the polar night transition. *Applied and Environmental Microbiology*, 78:4358–4366.
- Kraft B., Strous M., and Tegetmeyer H. E. (2011). Microbial nitrate respiration genes, enzymes and environmental distribution. *Journal of Biotechnology*, 155:104–117.
- Krause L., Diaz N. N., Goesmann A., Kelley S., Nattkemper T. W., Rohwer F., Edwards R. A., and Stoye J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, 36:2230–2239.
- Kumar S., Nei M., Dudley J., and Tamura K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinformatics*, 9:299–306.
- Kurosawa N., Sato S., Kawarabayashi Y., Imura S., and Naganuma T. (2010). Archaeal and bacterial community structures in the anoxic sediment of Antarctic meromictic lake Nurume-Ike. *Polar Science*, 4:421–429.
- La Scola B., Desnues C., Pagnier I., Robert C., Barrassi L., Fournous G., Merchat M., Suzan-Monti M., Forterre P., Koonin E. V., and Raoult D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature*, 455:100–104.
- Labrenz M., Collins M. D., Lawson P. A., Tindall B. J., Schumann P., and Hirsch P. (1999). *ii₃Roseovarius tolerans*ii** gen. nov., sp. nov., a budding bacterium with variable bacteriochlorophyll a production from hypersaline Ekho Lake. *International Journal of Systematic Bacteriology*, 49:137–147.

- Larsen A., Castberg T., Sandaa R.-A., Brussaard C. P., Egge J., Heldal M., Paulino A., Thyrhaug R., Hannen E. van, and Bratbak G. (2001). Population dynamics and diversity of phytoplankton, bacteria and viruses in a seawater enclosure. *Marine Ecology Progress Series*, 221:47–57.
- Larsen J. B., Larsen A., Bratbak G., and Sandaa R.-A. (2008). Phylogenetic analysis of members of the *Phycodnaviridae* virus family, using amplified fragments of the major capsid protein gene. *Applied and environmental microbiology*, 74:3048–3057.
- Laserson J., Jojic V., and Koller D. (2011). Genovo: *de novo* assembly for metagenomes. *Journal of Computational Biology*, 18:429–443.
- Lauro F. M., McDougald D., Thomas T., Williams T. J., Egan S., Rice S., DeMaere M. Z., Ting L., Ertan H., Johnson J., Ferriera S., Lapidus A., Anderson I. J., Kyrpides N., Munk A. C., Detter C., Han C. S., Brown M. V., Robb F. T., Kjelleberg S., and Cavicchioli R. (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences USA*, 106:15527–15533.
- Lauro F. M., DeMaere M. Z., Yau S., Brown M. V., Ng C., Wilkins D., Raftery M. J., Gibson J. A., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Thomas T., and Cavicchioli R. (2011). An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME Journal*, 5:879–895.
- Lavire C., Normand P., Alekhina I., Bulat S., Prieur D., Birrien J.-L., Fournier P., Hänni C., and Petit J.-R. (2006). Presence of Hydrogenophilus thermoluteolus DNA in accretion ice in the subglacial Lake Vostok, Antarctica, assessed using *rrs*, *cbb*, and *hox*. *Environmental Microbiology*, 8:2106–2114.
- Law P. (1959). The Vestfold Hills. *ANARE Reports*, 1:1–50.
- Laybourn-Parry J. The microbial loop in Antarctic lakes. In Howard-Willsiams C., Lyons W., and Hawes I., editors, *Ecosystem Dynamics in a Antarctic Ice-Free Landscapes*, pages 231–240. Rotterdam, 1997.
- Laybourn-Parry J. and Pearce D. A. (2007). The biodiversity and ecology of Antarctic lakes: models for evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 362:2273–2289.
- Laybourn-Parry J., Hofer J. S., and Sommaruga R. (2001). Viruses in the plankton of freshwater and saline Antarctic lakes. *Freshwater Biology*, 46:1279–1287.
- Laybourn-Parry J., Marshall W. A., and Marchant H. J. (2005). Flagellate nutritional versatility as a key to survival in two contrasting Antarctic saline lakes. *Freshwater Biology*, 50:830–838.
- Ley R. E., Turnbaugh P. J., Klein S., and Gordon J. I. (2006). Human gut microbes associated with obesity. *Nature*, 444:1022–1023.
- Li L., Stoeckert C. J., and Roos D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13:2178–2189.
- López-Bueno A., Tamares J., Velázquez D., Moya A., Quesada A., and Alcamí A. (2009). High diversity of the viral community from an Antarctic lake. *Science*, 858: 859–861.
- Lotka A. (1910). Contribution to the theory of periodic reaction. *Journal of Physical Chemistry*, 14:271–274.

- Lovelock J. E. and Maggs R. (1972). Atmospheric dimethyl sulphide and the natural sulphur cycle. *Nature*, 237:452–453.
- Madan N. J., Marshall W. A., and Laybourn-Parry J. (2005). Virus and microbial loop dynamics over an annual cycle in three contrasting Antarctic lakes. *Freshwater B*, 50:1291–1300.
- Man D., Wang W., Sabehi G., Aravind L., Post A. F., Massana R., Spudich E. N., Spudich J. L., and Béjà O. (2003). Diversification and spectral tuning in marine proteorhodopsins. *The EMBO Journal*, 22:1725–1731.
- Marcotte E. M. (2007). How do shotgun proteomics algorithms identify proteins? *Nature Biotechnology*, 25:755–757.
- Markowitz V. M., Ivanova N. N., Szeto E., Palaniappan K., Chu K., Dalevi D., Chen I.-M. A., Grechkin Y., Dubchak I., Anderson I. J., Lykidis A., Mavromatis K., Hugenholtz P., and Kyprides N. C. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Research*, 36:D534–8.
- Markowitz V. M., Chen I.-M. A., Chu K., Szeto E., Palaniappan K., Grechkin Y., Ratner A., Jacob B., Pati A., Huntemann M., Liolios K., Pagani I., Anderson I., Mavromatis K., Ivanova N. N., and Kyprides N. C. (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Research*, 40:D123–9.
- Martínez-Martínez J., Schroeder D. C., Larsen A., Bratbak G., and Wilson W. H. (2007). Molecular dynamics of *Emiliania huxleyi* and cooccurring viruses during two separate mesocosm studies. *Applied and Environmental Microbiology*, 73:554–562.
- Martínez Martínez J., Poulton N. J., Stepanauskas R., Sieracki M. E., and Wilson W. H. (2011). Targeted sorting of single virus-infected cells of the coccinolitophore *Emiliania huxleyi*. *PloS One*, 6:e22520.
- Matsuzaki M., Kubota K., Satoh T., Kunugi M., Ban S., and Imura S. (2006). Dimethyl sulfoxide-respiring bacteria in Suribati Ike, a hypersaline lake, in Antarctica and the marine environment. *Polar Bioscience*, 20:73–81.
- McHardy A. C., Martín H. G., Tsirigos A., Hugenholtz P., and Rigoutsos I. (2006). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4:63–72.
- Meinicke P., Asshauer K. P., and Lingner T. (2011). Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, 27:1618–1624.
- Meyer F., Paarmann D., D’Souza M., Olson R., Glass E., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., and Edwards R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386.
- Mikucki J. A. and Priscu J. C. (2007). Bacterial diversity associated with Blood Falls, a subglacial outflow from the Taylor Glacier, Antarctica. *Applied and Environmental Microbiology*, 73:4029–4039.
- Mikucki J. A., Pearson A., Johnston D. T., Turchyn A. V., Farquhar J., Schrag D. P., Anbar A. D., Priscu J. C., and Lee P. A. (2009). A contemporary microbially maintained subglacial ferrous “ocean”. *Science*, 324:397–400.

- Miyoshi T., Iwatsuki T., and Naganuma T. (2005). Phylogenetic characterization of 16S rRNA gene clones from deep-groundwater microorganisms that pass through 0.2-micrometer-pore-size filters. *Applied and environmental microbiology*, 71:1084–1088.
- Monier A., Claverie J.-M., and Ogata H. (2008). Taxonomic distribution of large DNA viruses in the sea. *Genome biology*, 9:R106.
- Monier A., Larsen J. B., Sandaa R.-A., Bratbak G., Claverie J.-M., and Ogata H. (2008). Marine mimivirus relatives are probably large algal viruses. *Virology Journal*, 5:12.
- Moran M. A. and Miller W. L. (2007). Resourceful heterotrophs make the most of light in the coastal ocean. *Nature Reviews Microbiology*, 5:792–800.
- Moran M. A., Belas R., Schell M., González J. M., Sun F., Sun S., Binder B. J., Edmonds J., Ye W., Orcutt B., Howard E. C., Meile C., Palefsky W., Goesmann A., Ren Q., Paulsen I., Ulrich L., Thompson L., Saunders E., and Buchan A. (2007). Ecological genomics of marine Roseobacters. *Applied and Environmental Microbiology*, 73: 4559–4569.
- Moran M. A., Reisch C. R., Kiene R. P., and Whitman W. B. (2012). Genomic insights into bacterial DMSP transformations. *Annual Review of Marine Science*, 4:523–542.
- Morris R. M., Nunn B. L., Frazer C., Goodlett D. R., Ting Y. S., and Rocap G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *The ISME Journal*, 4:673–685.
- Mosier A. C., Murray A. E., and Fritsen C. H. (2007). Microbiota within the perennial ice cover of Lake Vida, Antarctica. *FEMS Microbiology Ecology*, 59:274–288.
- Myers E. W., Sutton G. G., Delcher A. L., Dew I. M., Fasulo D. P., Flanigan M. J., Kravitz S. A., Mobarry C. M., Reinert K. H., Remington K. A., Anson E. L., Bolanos R. A., Chou H.-H., Jordan C. M., Halpern A. L., Lonardi S., Beasley E. M., Brandon R. C., Chen L., Dunn P. J., Lai Z., Liang Y., Nusskern D. R., Zhan M., Zhang Q., Zheng X., Rubin G. M., Adams M. D., and Venter J. C. (2000). A whole-genome assembly of *Drosophila*. *Science*, 287:2196–2204.
- Nagasaki K., Ando M., Imai I., Itakura S., and Ishida Y. (1994). Virus-like particles in *Heterosigma akashiwo*: a possible red tide disintegration mechanism. *Marine Biology*, 119:307–312.
- Nagasaki K., Shirai Y., Tomaru Y., Nishida K., and Pietrokovski S. (2005). Algal viruses with distinct intraspecies host specificities include identical intein elements. *Applied and Environmental Microbiology*, 71:3599–3607.
- Namiki T., Hachiya T., Tanaka H., and Sakakibara Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40:e155.
- Nelson M., Zhang Y., and Van Etten J. L. DNA methyltransferases and DNA site-specific endonucleases encoded by chlorella viruses. In Jost J. and Saluz H., editors, *DNA methylation: molecular biology and biological significance*, pages 186–211. Birkhäuser, Basel, 1993.
- Ng C. *A metaproteomic analysis of microbial communities in Ace Lake, Antarctica*. PhD thesis, University of New South Wales, 2010.

- Ng C., DeMaere M. Z., Williams T. J., Lauro F. M., Raftery M., Gibson J. A., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Thomas T., and Cavicchioli R. (2010). Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME Journal*, 4:1002–19.
- Noble R. T. and Fuhrman J. A. (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology*, 14:113–118.
- Noguchi H., Park J., and Takagi T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*, 34:5623–5630.
- Olson J., Steppe T., Litaker R., and Paerl H. (1998). N₂-fixing microbial consortia associated with the ice cover of Lake Bonney, Antarctica. *Microbial Ecology*, 36:231–238.
- Parks D. H. and Beiko R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26:715–721.
- Parks D. H., MacDonald N. J., and Beiko R. G. (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, 12:328.
- Patel A., Noble R. T., Steele J. A., Schwalbach M. S., Hewson I., and Fuhrman J. A. (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature Protocols*, 2:269–276.
- Pearce D. A. (2003). Bacterioplankton community structure in a maritime antarctic oligotrophic lake during a period of holomixis, as determined by denaturing gradient gel electrophoresis (DGGE) and fluorescence in situ hybridization (FISH). *Microbial Ecology*, 46:92–105.
- Pearce D. A. (2005). The structure and stability of the bacterioplankton community in Antarctic freshwater lakes, subject to extremely rapid environmental change. *FEMS Microbiology Ecology*, 53:61–72.
- Pearce D. A., Gast C. J., Lawley B., and Ellis-Evans J. C. (2003). Bacterioplankton community diversity in a maritime Antarctic lake, determined by culture-dependent and culture-independent techniques. *FEMS Microbiology Ecology*, 45:59–70.
- Pearce D. A., Gast C. J., van der Woodward K., and Newsham K. K. (2005). Significant changes in the bacterioplankton community structure of a maritime Antarctic freshwater lake following nutrient enrichment. *Microbiology*, 151:3237–3248.
- Pell J., Hintze A., Canino-Koning R., Howe A., Tiedje J. M., and Brown C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences USA*, 109:13272–13277.
- Peng Y., Leung H. C. M., Yiu S. M., and Chin F. Y. L. (2011). Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics*, 27:i94–i101.
- Pickard J., Adamson D. A., and Heath C. W. (1986). The evolution of Watts Lake, Vestfold Hills, East Antarctica, from marine inlet to freshwater lake. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 53:271–288.
- Pignatelli M. and Moya A. (2011). Evaluating the fidelity of *de novo* short read metagenomic assembly using simulated data. *PloS One*, 6:e19984.

- Poretsky R. S., Sun S., Mou X., and Moran M. A. (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environmental Microbiology*, 12:616–627.
- Priscu J. C. (1999). Geomicrobiology of subglacial ice above Lake Vostok, Antarctica. *Science*, 286:2141–2144.
- Proctor L. M. and Fuhrman J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature*, 343:60–62.
- Proctor L. M. and Fuhrman J. A. (1992). Mortality of marine bacteria in response to enrichments of the virus size fraction from seawater. *Marine Ecology Progress Series*, 87:283–293.
- Purdy K., Nedwell D., and Embley T. (2003). Analysis of the sulfate-reducing bacterial and methanogenic archaeal populations in contrasting Antarctic sediments. *Applied and Environmental Microbiology*, 69:3181–3191.
- Quayle W. C., Peck L. S., Peat H., Ellis-Evans J., and Harrigan P. R. (2002). Extreme responses to climate change in Antarctic lakes. *Science*, 295:645.
- Raina J.-B., Dinsdale E. A., Willis B. L., and Bourne D. G. (2010). Do the organic sulfur compounds DMSP and DMS drive coral microbial associations? *Trends in Microbiology*, 18:101–108.
- Ram R. J., Verberkmoes N. C., Thelen M. P., Tyson G. W., Baker B. J., Blake II R. C., Shah M., Hettich R. L., and Banfield J. F. (2005). Community proteomics of a natural microbial biofilm. *Science*, 308:1915–1920.
- Rankin L. M., Gibson J. A., Franzmann P. D., and Burton H. R. (1999). The chemical stratification and microbial communities of Ace Lake, Antarctica: a review of the characteristics of a marine-derived meromictic lake. *Polarforschung*, 66:33–52.
- Redfield A., Ketchum B., and Richards F. The influence of organisms on the composition of sea-water. In Hill M., editor, *The Sea*, pages 26–77. 1963. URL <http://www.vliz.be/imis/imis.php?refid=28944>.
- Reisch C. R., Moran M. A., and Whitman W. B. (2011). Bacterial catabolism of dimethylsulfoniopropionate (DMSP). *Frontiers in Microbiology*, 2:1–12.
- Rivière D., Desvignes V., Pelletier E., Chaussonnerie S., Guermazi S., Weissenbach J., Li T., Camacho P., and Sghir A. (2009). Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *The ISME Journal*, 3: 700–714.
- Roberts N. and Burton H. R. (1993). Sampling volatile organics from a meromictic Antarctic lake. *Polar Biology*, 13:359–361.
- Roberts N., Burton H. R., and Pitson G. (1993). Volatile organic compounds from Organic Lake, an Antarctic, hypersaline, meromictic lake. *Antarctic Science*, 5:361–366.
- Rodríguez-Brito B., Li L., Wegley L., Furlan M., Angly F. E., Breitbart M., Buchanan J., Desnues C., Dinsdale E. A., Edwards R. A., Felts B., Haynes M., Liu H., Lipson D., Mahaffy J., Martin-Cuadrado A. B., Mira A., Nulton J., Pašić L., Rayhawk S., Rodríguez-Mueller J., Rodríguez-Valera F., Salamon P., Srinagesh S., Thingstad

- T. F., Tran T., Thurber R. V., Willner D., Youle M., and Rohwer F. (2010). Viral and microbial community dynamics in four aquatic environments. *The ISME Journal*, 4: 739–751.
- Rosen G. L., Polikar R., Caseiro D. A., Essinger S. D., and Sokhansanj B. A. (2011). Discovering the unknown: improving detection of novel species and genera from short reads. *Journal of Biomedicine and Biotechnology*, page 495849.
- Röske K., Sachse R., Scheerer C., and Röske I. (2012). Microbial diversity and composition of the sediment in the drinking water reservoir Saidenbach (Saxonia, Germany). *Systematic and Applied Microbiology*, 35:35–44.
- Roux S., Faubladier M., Mahul A., Paulhe N., Bernard A., Debroas D., and Enault F. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27: 3074–3075.
- Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S. J., Yooseph S., Wu D., Eisen J. A., Hoffman J. M., Remington K., Beeson K. Y., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter J. E., Li K., Kravitz S., Heidelberg J. F., Utterback T., Rogers Y.-H., Falcón L. I., Souza V., Bonilla-Rosso G., Eguiarte L. E., Karl D. M., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari M. R., Strausberg R. L., Nealson K., Friedman R., Frazier M. E., and Venter J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5:e77.
- Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M.-A., and Barrell B. G. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, 16:944–945.
- Salzberg S. L., Phillippy A. M., Zimin A., Puiu D., Magoc T., Koren S., Treangen T. J., Schatz M. C., Delcher A. L., Roberts M., Marcais G., Pop M., and Yorke J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22:557–567.
- Samsudin A. A., Evans P. N., Wright A.-D. G., and Al Jassim R. (2011). Molecular diversity of the foregut bacteria community in the dromedary camel (*Camelus dromedarius*). *Environmental Microbiology*, 13:3024–3035.
- Sandaas R.-A., Heldal M., Castberg T., Thyrhaug R., and Bratbak G. (2001). Isolation and characterization of two viruses with large genome size infecting *Chrysochromulina ericina* (*Prymnesiophyceae*) and *Pyramimonas orientalis* (*Prasinophyceae*). *Virology*, 290:272–280.
- Saunders N. F. W., Ng C., Raftery M., Guilhaus M., Goodchild A., and Cavicchioli R. (2006). Proteomic and computational analysis of secreted proteins with type I signal peptides from the Antarctic archaeon *Methanococcoides burtonii*. *Journal of Proteome Research*, 5:2457–2464.
- Schäfer H., Myronova N., and Boden R. (2010). Microbial degradation of dimethyl-sulphide and related C₁-sulphur compounds: organisms and pathways controlling fluxes of sulphur in the biosphere. *Journal of Experimental Botany*, 61: 315–334.
- Schatz M. C., Langmead B., and Salzberg S. L. (2010). Cloud computing and the DNA data race. *Nature Biotechnology*, 28:691–693.

- Schiaffino M. R., Unrein F., Gasol J. M., Farias M. E., Estevez C., Balagué V., and Izaguirre I. (2009). Comparative analysis of bacterioplankton assemblages from maritime Antarctic freshwater lakes with contrasting trophic status. *Polar Biology*, 32: 923–936.
- Schmidtova J., Hallam S. J., and Baldwin S. A. (2009). Phylogenetic diversity of transition and anoxic zone bacterial communities within a near-shore anoxic basin: Nitinat Lake. *Environmental Microbiology*, 11:3233–3251.
- Scholz M. B., Lo C.-C., and Chain P. S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*, 23:9–15.
- Scott K. M., Sievert S. M., Abril F. N., Ball L. a., Barrett C. J., Blake R. a., Boller A. J., Chain P. S. G., Clark J. a., Davis C. R., Detter C., Do K. F., Dobrinski K. P., Faza B. I., Fitzpatrick K. a., Freyermuth S. K., Harmer T. L., Hauser L. J., Hügler M., Kerfeld C. a., Klotz M. G., Kong W. W., Land M., Lapidus A., Larimer F. W., Longo D. L., Lucas S., Malfatti S. a., Massey S. E., Martin D. D., McCuddin Z., Meyer F., Moore J. L., Ocampo L. H., Paul J. H., Paulsen I. T., Reep D. K., Ren Q., Ross R. L., Sato P. Y., Thomas P., Tinkham L. E., and Zeruth G. T. (2006). The genome of deep-sea vent chemolithoautotroph *Thiomicrospira crunogena* XCL-2. *PLoS Biology*, 4:e383.
- Sharma A. K., Zhaxybayeva O., Papke R. T., and Doolittle W. F. (2008). Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environmental Microbiology*, 10:1039–1056.
- Sharma A. K., Sommerfeld K., Bullerjahn G. S., Matteson A. R., Wilhelm S. W., Jezbera J., Brandt U., Doolittle W. F., and Hahn M. W. (2009). Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater *Actinobacteria*. *The ISME Journal*, 3:726–737.
- Sharma V. K., Kumar N., Prakash T., and Taylor T. D. (2012). Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. *PloS One*, 7:e34030.
- Siegert M. J., Ellis-Evans J. C., Tranter M., Mayer C., Petit J.-R., Salamatian A., and Priscu J. C. (2001). Physical, chemical and biological processes in Lake Vostok and other Antarctic subglacial lakes. *Nature*, 414:603–609.
- Sievert S. M., Scott K. M., Klotz M. G., Chain P. S., Hauser L. J., Hemp J., Hügler M., Land M., Lapidus A., Larimer F. W., Lucas S., Malfatti S. A., Meyer F., Paulsen I. T., Ren Q., and Simon J. (2008). Genome of the epsilonproteobacterial chemolithoautotroph *Sulfurimonas denitrificans*. *Applied and Environmental Microbiology*, 74:1145–1156.
- Sowell S. M., Wilhelm L. J., Norbeck A. D., Lipton M. S., Nicora C. D., Barofsky D. F., Carlson C. A., Smith R. D., and Giovanonni S. J. (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *The ISME Journal*, 3:93–105.
- Su X., Xu J., and Ning K. (2012). Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Systems Biology*, 6:S16.
- Sun S., Chen J., Li W., Altintas I., Lin A., Peltier S., Stocks K., Allen E. E., Ellisman M., Grethe J., and Wooley J. (2011). Community cyberinfrastructure for Advanced

- Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Research*, 39:D546–551.
- Suttle C. A. (2005). Viruses in the sea. *Nature*, 437:356–361.
- Suttle C. A., Chan A. M., and Cottrell M. T. (1990). Infection of phytoplankton by viruses and reduction of primary productivity. *Nature*, 347:467–469.
- Tajima K., Aminov R. I., Nagamine T., Ogata K., Nakamura M., Matsui H., and Benno Y. (1999). Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries. *FEMS Microbiology Ecology*, 29:159–169.
- Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28:2731–2739.
- Tang Y.-Q., Ji P., Hayashi J., Koike Y., Wu X.-L., and Kida K. (2011). Characteristic microbial community of a dry thermophilic methanogenic digester: its long-term stability and change with feeding. *Applied Microbiology and Biotechnology*, 91:1447–1461.
- Tatusov R. L., Koonin E. V., and Lipman D. J. (1997). A genomic perspective on protein families. *Science*, 278:631–637.
- Tatusov R. L., Fedorova N. D., Jackson J. D., Jacobs A. R., Kiryutin B., Koonin E. V., Krylov D. M., Mazumder R., Mekhedov S. L., Nikolskaya A. N., Rao B. S., Smirnov S., Sverdlov A. V., Vasudevan S., Wolf Y. I., Yin J. J., and Natale D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.
- Teeling H. and Glöckner F. O. (2012). Current opportunities and challenges in microbial metagenome analysis - a bioinformatic perspective. *Briefings in Bioinformatics*, 13 (6):728–742.
- Teeling H., Waldmann J., Lombardot T., Bauer M., and Glöckner F. O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 5:163.
- Teeling H., Fuchs B. M., Becher D., Klockow C., Gardebrecht A., Bennke C. M., Kassabgy M., Huang S., Mann A. J., Waldmann J., Weber M., Klindworth A., Otto A., Lange J., Bernhardt J., Reinsch C., Hecker M., Peplies J., Bockelmann F. D., Callies U., Gerdts G., Wichels A., Wiltshire K. H., Glöckner F. O., Schweder T., and Amann R. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science*, 336:608–611.
- Temperton B. and Giovannoni S. J. (2012). Metagenomics: microbial diversity through a scratched lens. *Current opinion in microbiology*, 15:605–12.
- Thakur R. S., Bandopadhyay R., Chaudhary B., and Chatterjee S. (2012). Now and next-generation sequencing techniques: future of sequence analysis using cloud computing. *Frontiers in Genetics*, 3:280.
- Thomsen H. A. (2007). Ultrastructural studies of the flagellate and cyst stages of *Pseudopedinella tricostata* (*Pedinellales*, *Chrysophyceae*). *British Phycological Journal*, pages 37–41.

- Todd J. D., Rogers R., Li Y. G., Wexler M., Bond P. L., Sun L., Curson A. R., Malin G., Steinke M., and Johnston A. W. (2007). Structural and regulatory genes required to make the gas dimethyl sulfide in bacteria. *Science*, 315:666–669.
- Todd J. D., Curson A. R., Dupont C. L., Nicholson P., and Johnston A. W. (2009). The β -dddP β gene, encoding a novel enzyme that converts dimethylsulfoniopropionate into dimethyl sulfide, is widespread in ocean metagenomes and marine bacteria and also occurs in some Ascomycete fungi. *Environmental Microbiology*, 11:1376–1385.
- Todd J. D., Curson A. R., Nikolaïdou-Katsaraidou N., Brearley C. A., Watmough N. J., Chan Y., Page P. C., Sun L., and Johnston A. W. (2010). Molecular dissection of bacterial acrylate catabolism - unexpected links with dimethylsulfoniopropionate catabolism and dimethyl sulfide production. *Environmental Microbiology*, 12:327–343.
- Todd J. D., Curson A. R., Kirkwood M., Sullivan M. J., Green R. T., and Johnston A. W. (2011). DddQ, a novel, cupin-containing, dimethylsulfoniopropionate lyase in marine roseobacters and in uncultured marine bacteria. *Environmental Microbiology*, 13:427–438.
- Todd J. D., Kirkwood M., Newton-Payne S., and Johnston A. W. (2012). DddW, a third DMSP lyase in a model *Roseobacter* marine bacterium, *Ruegeria pomeroyi* DSS-3. *The ISME Journal*, 6:223–226.
- Torrice M. Viral ecology hit by filter shortage, 2009. URL <http://news.sciencemag.org/scienceinsider/2009/10/viral-ecology-r.html>.
- Tyson G. W., Chapman J., Hugenholtz P., Allen E. E., Ram R. J., Richardson P. M., Solovyev V. V., Rubin E. M., Rokhsar D. S., and Banfield J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37–43.
- Unrein F., Izaguirre I., Massana R., Balagué V., and Gasol J. M. (2005). Nanoplankton assemblages in maritime Antarctic lakes: characterisation and molecular fingerprinting comparison. *Aquatic Microbial Ecology*, 40:269–282.
- Hoff J. van den and Franzmann P. D. (1986). A choanoflagellate in a hypersaline Antarctic lake. *Polar Biology*, 6:71–73.
- Van Etten J. L., Lane L. C., and Meints R. H. (1991). Viruses and viruslike particles of eukaryotic algae. *Microbiological Reviews*, 55(4):586–620.
- Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D. B., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W., Fouts D. E., Levy S., Knap A. H., Lomas M. W., Nealson K., White O., Peterson J., Hoffman J. M., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y.-H., and Smith H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66–74.
- Villaescusa J. A., Casamayor E. O., Rochera C., Velázquez D., Álvaro C., Quesada A., and Camacho A. (2010). A close link between bacterial community composition and environmental heterogeneity in maritime Antarctic lakes. *International Microbiology*, 13:67–77.
- Volterra V. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. R. Accad. Naz. dei Lincei*, 2:31–113.

- Mering C.von, Hugenholtz P., Raes J., Tringe S. G., Doerks T., Jensen L., Ward N., and Bork P. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315:1126–1130.
- Voytek M. A., Priscu J. C., and Ward B. B. (1999). The distribution and relative abundance of ammonia-oxidizing bacteria in lakes of the McMurdo Dry Valley, Antarctica. *Hydrobiologia*, 401:113–130.
- Wagner-Döbler I. and Biebl H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annual Review of Microbiology*, 60:255–280.
- Wang Q., Garrity G. M., Tiedje J. M., and Cole J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73:5261–5267.
- Ward B. B. and Priscu J. C. (1997). Detection and characterization of denitrifying bacteria from a permanently ice-covered Antarctic lake. *Hydrobiologia*, pages 57–68.
- Ward B. B., Granger J., Maldonado M., Casciotti K., Harris S., and Wells M. (2005). Denitrification in the hypolimnion of permanently ice-covered Lake Bonney, Antarctica. *Aquatic Microbial Ecology*, 38:295–307.
- Warnecke F. and Hugenholtz P. (2007). Building on basic metagenomics with complementary technologies. *Genome Biology*, 8:231.
- Wen K., Ortmann A. C., and Suttle C. A. (2004). Accurate estimation of viral abundance by epifluorescence microscopy. *Applied and Environmental Microbiology*, 70: 3862–3867.
- Wilkening J., Wilke A., Desai N., and Meyer F. (2009). Using clouds for metagenomics: a case study. *IEEE International Conference*, pages 1–6.
- Wilkins D., Lauro F. M., Williams T. J., DeMaere M. Z., Brown M. V., Hoffman J. M., Andrews-Pfannkoch C., Mcquaid J. B., Riddle M. J., Rintoul S. R., and Cavicchioli R. (2012). Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environmental Microbiology*, page (in press).
- Williams T. J., Wilkins D., Long E., Evans F., DeMaere M. Z., Raftery M. J., and Cavicchioli R. (2012). The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environmental Microbiology*, page (in press).
- Wilson W. H., Tarhan G. A., Schroeder D., Cox M., Oke J., and Malin G. (2002). Isolation of viruses responsible for the demise of an *Emiliania huxleyi* bloom in the English Channel. *Journal of the Marine Biological Association of the United Kingdom*, 82:369–377.
- Wommack K. E., Bhavsar J., Polson S. W., Chen J., Dumas M., Srinivasiah S., Furman M., Jamindar S., and Nasko D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, 6:427–439.
- Wu S., Zhu Z., Fu L., Niu B., and Li W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics*, 12:444.
- Wu Y.-W. and Ye Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18:523–534.

- Xie C., Mao X., Huang J., Ding Y., Wu J., Dong S., Kong L., Gao G., Li C., and Wei L. (2011). KOBAS 2.0: a web server for the annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, 39:W316–322.
- Xing P., Hahn M. W., and Wu Q. L. (2009). Low taxon richness of bacterioplankton in high-altitude lakes of the Eastern Tibetan Plateau, with a predominance of *Bacteroidetes* and *Synechococcus* spp. *Applied and Environmental Microbiology*, 75:7017–7025.
- Yamamoto M. and Takai K. (2011). Sulfur metabolisms in epsilon- and gamma-*Proteobacteria* in deep-sea hydrothermal fields. *Frontiers in Microbiology*, 2: 1–8.
- Yamane K., Hattori Y., Ohtagaki H., and Fujiwara K. (2011). Microbial diversity with dominance of 16S rRNA gene sequences with high GC contents at 74 and 98 °C subsurface crude oil deposits in Japan. *FEMS Microbiology Ecology*, 76:220–235.
- Yanagibayashi M., Nogi Y., Li L., and Kato C. (1999). Changes in the microbial community in Japan Trench sediment from a depth of 6292 m during cultivation without decompression. *FEMS Microbiology Letters*, 170:271–279.
- Yau S., Lauro F. M., DeMaere M. Z., Brown M. V., Thomas T., Raftery M. J., Andrews-Pfannkoch C., Lewis M., Hoffman J. M., Gibson J. A., and Cavicchioli R. (2011). Virophage control of antarctic algal hostvirus dynamics. *Proceedings of the National Academy of Sciences USA*, 108:6163–6168.
- Yilmaz P., Iversen M. H., Hankeln W., Kottmann R., Quast C., and Glöckner F. O. (2012). Ecological structuring of bacterial and archaeal taxa in surface ocean waters. *FEMS Microbiology Ecology*, 81:373–385.
- Yilmaz S., Allgaier M., and Hugenholtz P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods*, 7:943–944.
- Zwart D., Bird M., Stone J., and Lambeck K. (1998). Holocene sea-level change and ice-sheet history in the Vestfold Hills, East Antarctica. *Earth and Planetary Science Letters*, 155:131–145.
- Zybailov B., Mosley A. L., Sardiu M. E., Coleman M. K., Florens L., and Washburn M. P. (2006). Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae* research articles. *Journal of Proteome Research*, 5: 2339–2347.

Appendix A

PCR-based studies of Antarctic Lakes

Table A.1: Studies of Antarctic Lakes that have made use of PCR amplification and sequencing of marker genes. The list presented here has attempted to be comprehensive but some studies some may have been inadvertently missed. DV, McMurdo Dry Valleys; VH, Vestfold Hills; Syo, Syowa Oasis; SI, Signy Island; AP, Antarctic Peninsula; KGI, King George Island; LI, Livingston Island; CFB, *Cytophaga*, *Flavobacteria*, *Bacteroidetes* group; *Alpha-*, *Alphaproteobacteria*; *Beta-*, *Betaproteobacteria*; *Gamma-*, *Gammaproteobacteria*; *Delta-*, *Deltaproteobacteria*; *Epsilon-*, *Epsilonproteobacteria*; *Actino-*, *Actinobacteria*; *Cyano-*, *Cyanobacteria*; SRB, sulphate-reducing bacteria; FISH, fluorescence *in situ* hybridisation; FAME, fatty-acid methyl ester.

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|---------------------------------------|--|---|---|------------------------------|--|----------------------------------|
| Lakes Bonney, Hoare, Joyce, Vanda, DV | Fresh to hypersaline, permanently ice-covered | 16S, <i>amoA</i> libraries | <i>Beta-</i> , <i>Gamma-</i> | Ammonia | Nitrifying bacterial <i>amoA</i> detected in all lakes. In meromictic lakes, the population of <i>Beta</i> and <i>Gamma</i> - vertically stratified. Majority of nitrifying bacteria were <i>Beta</i> -. | (Voytek <i>et al.</i> , 1999) |
| Lake Bonney, DV | Hypersaline, meromictic, permanently ice-covered, separated into east and west lobes | <i>nifH</i> library of ice aggregate material and microbial mats, nitrogenase assay | <i>Cyano-</i> , <i>Gamma-</i> , <i>Alpha-</i> , <i>Delta-</i> | N ₂ fixation | Nitrogenase activity low compared to temperate environments. Heterotrophs responsible for 10–30% of nitrogenase activity. Heterotrophs likely microaerophilic. | (Olson <i>et al.</i> , 1998) |
| Lake Bonney, cyanobacterial mats, DV | Hypersaline, meromictic, permanently ice-covered, separated into east and west lobes | 16S library of ice sediments, hybridisation of probes | <i>Cyano-</i> , <i>Acidobacterium/Holophaga</i> , green non-sulphur bacteria | Phototrophy and heterotrophy | Probes designed from 16S clone library of bacteria in the sediment in the ice matched that of the surrounding mats. | (Gordon <i>et al.</i> , 2000) |
| Lake Bonney, DV | Hypersaline, meromictic, permanently ice-covered, separated into east and west lobes | 18S libraries of watercolumn | <i>Cryptophyta</i> , <i>Chlorophyta</i> , <i>Stramenopiles</i> , <i>Haptophyta</i> , <i>Choanoflagellida</i> , <i>Alveolata</i> , fungi, ciliates | Photosynthesis | Population vertically stratified. Cryptophytes dominant in the shallow water and haptophytes in the mid-depths and chlorophytes in the deeper waters. Stramenopiles replaced haptophytes during polar night. | (Bielewicz <i>et al.</i> , 2011) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|----------------------------------|--|---|--|---------------------------|--|----------------------------------|
| Ekho, Organic, Deep Lakes, VH | Hypersaline. Ekho and Organic: meromictic and ice-covered ~9 months of the year. Deep: holomictic, never freezes. | 16S libraries of sediment | Organic: <i>Cyano-/chloroplasts, CFB, Gamma-, Alpha-, Halobacteriales, Actino-.</i> Ekho: <i>Cyano-/chloroplasts, CFB, Firmicutes, Alpha-, Gamma-, Verrucomicrobiales, Spirochaetales.</i> Deep: <i>Halobacteriales, Gamma-</i> | Heterotrophy | No phylotypes common to all samples. Distribution of bacterial classes similar between Ekho and Organic with <i>Roseovarius</i> common to both. <i>Marinobacter</i> and <i>Halomonas</i> common to Organic and Deep. | (Bowman <i>et al.</i> , 2000a) |
| Lake Vida, DV | Hypersaline, meromictic, permanently ice-covered | 16S, 18S DGGE, 16S library of ice core | 16S: <i>Actino-, CFB, Gamma-, Cyano-, OD1, TM7, Firmicutes, Planctomycetales.</i> 18S: <i>Chlorophyta, fungi, Bacillariophyta, Apicomplexa, Cercozoa, Chrysophyceae, Ciliophora</i> | Phototrophy, heterotrophy | Cell density highest at the surface. Phylogeny shows <i>Marinobacter</i> related to Lake Bonney isolate and bacterial sequences are similar to marine and polar organisms. | (Mosier <i>et al.</i> , 2007) |
| Suribati-Ike, Syo | Hypersaline, meromictic, sulphidic anoxic bottom waters. | 16S libraries of halocline | <i>Marinobacter, Halomonas, Pseudomonas, Halocella</i> | Heterotrophy | <i>Marinobacter</i> isolates capable of DMSO-respiration were relatives of those detected in lake water. Bacteria from the water column unable to respire nitrate. | (Matsuzaki <i>et al.</i> , 2006) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|---|--|--|---|---|---|--------------------------------|
| Clear, Pendant, Scale, Ace, Burton Lakes, Tay-naya Bay, VH | Saline, meromictic lakes, high levels of sulphides (120–>250 mmol kg ⁻¹) | 16S libraries of anoxic sediment | Bacteria: <i>Firmicutes</i> , <i>Cyanobacteria</i> /chloroplasts, CFB, <i>Delta-</i> , <i>Alpha-</i> , <i>Planctomycetes</i> , <i>Gamma-</i> , green non-sulphur bacteria, <i>Chlamydiales</i> , <i>Verrucomicrobia</i> , <i>Actinobacteria</i> . Eucarya: 2.5% of clones. Archaea: <i>Methanosaerina barkerii</i> , unknown <i>Euryarchaeota</i> group equidistant from <i>Thermoplasma</i> , <i>Methanomicrobiales</i> , <i>Halobacteriales</i> | Sulphate reduction, methanogenesis, phototrophy, aerobic heterotrophy | Lakes with similar physico-chemical and limnological traits had more similar microbial communities. | (Bowman <i>et al.</i> , 2000b) |
| Lake Fryxell, DV | Brackish, meromictic, permanently ice-covered | <i>pufM</i> libraries, DGGE, RT-PCR of <i>pufM</i> transcripts | <i>Alpha-, Beta-</i> related to purple non-sulphur bacteria and aerobic anoxygenic phototrophs | Anoxygenic photosynthesis | Vertical stratification of the community down the water column. Purple and green sulphur bacteria not detected despite the high sulphide. <i>pufM</i> transcripts only found below 9 m even though <i>pufM</i> genes found throughout the water column. | (Karr <i>et al.</i> , 2003) |
| Lake Fryxell, DV | Brackish, meromictic, permanently ice-covered | 16S DGGE of water column | <i>Methanoculleus</i> , <i>Methanosaerina</i> , unclassified <i>Euryarchaeota</i> , <i>Methanosaericinales</i> -group and marine benthic group C-like <i>Crenarchaeota</i> | Hydrogenotrophic methanogenesis, potential anoxic methanotrophy | Diverse population of methanogenic <i>Euryarchaeota</i> , unclassified <i>Euryarchaeota</i> and divergent <i>Crenarchaeota</i> detected in sediments and water column. | (Karr <i>et al.</i> , 2006) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|-----------------|--------------------|--------------------------------|--|---------------|--|---------------------------------|
| Nurume-Ike, Syo | Saline, meromictic | 16S library of anoxic sediment | Archaea: marine benthic group and unclassified <i>Euryarchaeota</i> . Bacteria: <i>Alpha-</i> , <i>Delta-</i> , <i>Planctomycetes</i> , <i>Cyano-</i> /chloroplast, <i>Gamma-</i> , <i>Actino-</i> , CFB, <i>Verrucomicrobia</i> and <i>Spirochaetes</i> | Heterotrophy | Distribution of bacterial classes similar to lake sediment in VH except <i>Alpha</i> -relatively overrepresented and <i>Firmicutes</i> underrepresented. | (Kurosawa <i>et al.</i> , 2010) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|----------------------------------|---|--|--|---|--|------------------------------|
| Heywood Lake, Shallow Bay, SI | Heywood: ice covered for ~9 months of the year, eutrophic due to organic inputs from seals, separated into two basins by shallow inlet. Shallow: coastal marine, ice covered during winter | Archaeal 16S, universal 16S libraries of anoxic sediment. Northern blots with methanogenic archaeal probes | Heywood blots: <i>Methanomicrobiales</i> , <i>Methanogenium</i> , <i>Methanosarcinales</i> , <i>Methanosaeta</i> . Shallow blots: <i>Methanosarcinales</i> , <i>Methanomicrobiales</i> , <i>Methanococcales</i> . Heywood Archaea: <i>Methanosaeta</i> , <i>Methanogenium</i> . Shallow Archaea: <i>Methanogenium</i> , <i>Methanolobus</i> , <i>Methanococcales</i> . Heywood SRB: <i>Desulfovibrio</i> , <i>Desulfotalea</i> / <i>Desulforhopalus</i> , <i>Desulfobulbus</i> , <i>Desulfobacteriaceae</i> . Shallow Bay SRB: <i>Desulfotalea</i> / <i>Desulforhopalus</i> , <i>Desulfobacterium</i> , <i>Desulfobulbus</i> , <i>Desulfobacteriaceae</i> | Acetoclastic and hydrogenotrophic methanogenesis, sulphur and metal oxidation, sulphate reduction | Methanogenesis and sulphate reduction detected at both sites. Diversity of methanogenic <i>Archaea</i> extremely low. Methanogenic archaea 34% and 0.2% of community in Heywood Lake and Shallow Bay respectively. SRB 0.9% and 14.7% of community in Heywood Lake and Shallow Bay respectively. | (Purdy <i>et al.</i> , 2003) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|---------------------------------|---|---|--|---|---|-------------------------------|
| Sombre Lake, SI | Freshwater, ice-covered for ~9 months of the year, oligotrophic, N and P limited | 16S libraries, DGGE, 16S libraries, FAME analysis of isolates, FISH of water column profile | 16S of isolates: <i>Beta-, Firmicutes, Actinobacteria, Alph-, Gamma-</i> . FAME: <i>Firmicutes, Actino-, Gamma-, Beta-, Alpha-</i> . Clones: <i>Actino-, CFB, Beta-, Alpha-, Spirochaetales, Delta-, Gamma-, Verrucomicrobia.</i> FISH: <i>Beta-, CFB, Alpha-, Gamma-</i> . DGGE: <i>Actinobacteria, CFB, Beta-</i> | Heterotrophic mainly respiratory metabolism | Relative abundances shown by clone libraries and FISH the same. Few genera common to culture-dependent and independent techniques. 16S isolate library and 16S clone library were significantly different. 16S clone library covers the largest spread of phyla but is missing <i>Firmicutes</i> . Overall <i>Beta-</i> . | (Pearce, 2003) |
| Moss Lake, SI | Freshwater, ice-covered for ~9 months of the year, oligotrophic, N and P limited | 16S DGGE and FISH of water column | <i>Beta-, CFB, Alpha-, Gamma-, Actino-, Cyano-</i> . <1% of cells hybridised with archaeal FISH probe | Heterotrophy, mainly respiratory metabolism | Very little vertical stratification of population. 16S sequences similar to temperate and cold aquatic systems. | (Pearce <i>et al.</i> , 2003) |
| Moss, Sombre, Heywood Lakes, SI | Freshwater, ice-covered for ~9 months of the year, oligotrophic to eutrophic status | 16S DGGE of water column profile over the winter to summer transition | Not determined | Not determined | Lakes were physically and chemically stratified in winter, mixed in summer. Variation in bacterial community structure correlated with lake chemistry. Bacterial community still unstable during holomixis. | (Pearce, 2005) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|--|--|---|--|---|---|-----------------------------------|
| Heywood Lake, SI | Freshwater, ice-covered for ~9 months of the year, eutrophic due to organic inputs from seals, separated into two basins by shallow inlet. | 16S libraries, DGGE, 16S libraries, FAME analysis of isolates, FISH of water column profile | 16S clones: <i>Beta</i> -, <i>Alpha</i> -, <i>Actino</i> -. FAME: <i>Actino</i> -, <i>Firmicutes</i> , <i>Gamma</i> -, <i>Alpha</i> -. FISH: <i>Beta</i> -, CFB, <i>Gamma</i> -, <i>Alpha</i> -. DGGE: <i>Actino</i> -, CFB, Gram-positives, <i>Beta</i> -. | Heterotrophy, mainly respiratory metabolisms, phototrophy | Clone library coverage 71.7%. Similar genera to Moss and Sombre Lakes. <i>Actino</i> - and marine <i>Alpha</i> - enriched compared to oligotrophic lakes while <i>Cyano</i> - underrepresented. Species evenness is higher than Sombre or Moss Lakes. | (Pearce <i>et al.</i> , 2005) |
| Lakes Boeckella, Esperanza, Flora, Encantado, Chico, Pingüi, Hope Bay, AP. Lakes L, M, W, Z, KGI | Freshwater, oligotrophic except Pingüi and Boeckella which were eutrophic and mesotrophic respectively | 18S DGGE of surface water (20–3 µm) | <i>Chrysophyta</i> , <i>Chlorophyta</i> , <i>Dictyochophyceae</i> , <i>Bacillariophyceae</i> , <i>Cercozoa</i> | Photosynthesis | Molecular surveys showed a greater level of diversity exists than can be determined by light microscopy. Lake communities varied depending on trophic status. Lakes in both regions shared bands belonging to <i>Chrysophyta</i> although they were 220 km apart. <i>Dictyochophyceae</i> and <i>Cercozoa</i> restricted to oligotrophic lakes. | (Unrein <i>et al.</i> , 2005) |
| Lakes Boeckella, Esperanza, Flora, Encantado, Chico, Pingüi, Hope Bay, AP. Lakes W and Z, KGI. | Freshwater, oligotrophic except Pingüi and Boeckella which were eutrophic and mesotrophic respectively. | 16S DGGE of surface water (20–3 µm) | CFB, <i>Actino</i> -, <i>Beta</i> -, <i>Cyano</i> - | Heterotrophy, photosynthesis | Cluster analysis showed Lake communities from the Hope Bay formed one group while Lakes Chico, Pingüi and Boeckella formed another subgroup with KGI lakes. 63.7% of variance is explained by axis 1 and 2 of Canonical Correspondence Analysis (40.4% phosphate, dissolved inorganic nitrogen and pH; 23.3% dissolved inorganic nitrogen). Temporal variation is not as pronounced as differences due to trophic status. | (Schiaffino <i>et al.</i> , 2009) |

Continued on next page

Table A.1 – *Continued from previous page*

| Site | Environment | Techniques | Organisms | Key processes | Notes | Reference |
|--|---|---|--|------------------------------------|---|------------------------------------|
| Lakes Limnopolar, Midge, Chester, Chica, Turbio, Somero, Refugio, LI | Fresh to saline, all oligotrophic except for Refugio, which was eutrophic | 16S DGGE of from surface water | CFB, <i>Alpha-</i> | Heterotrophy, phototrophy | Cluster analysis showed deep lakes of the plateau grouped together while Somero and Refugio were separate groups. Over 90% of variance was explained by chemical parameters related to trophic status and salinity. | (Villaescusa <i>et al.</i> , 2010) |
| Lake Vostok | Largest subglacial lake, isolated from surface 420,000 years | 16S library of accretion ice core from 3,590 m | <i>Alpha-, Beta-, Actinomycetes</i> | Potential heterotrophy | No <i>Archaea</i> were amplified with archaeal primers. No biological incorporation of selected substrates | (Priscu, 1999) |
| Lake Vostok | Largest subglacial lake, isolated from surface 420,000 years | 16S library of accretion ice core from 3,590 and 3,603 m and isolation of bacteria | <i>Alpha-, Beta-, Firmicutes, Actino-, CFB</i> | Potential heterotrophy | Bacteria appear related to isolates from similarly cold environments. No <i>Archaea</i> were amplified using archaeal primers. | (Christner <i>et al.</i> , 2001) |
| Lake Vostok | Largest subglacial lake, isolated from surface 420,000 years | 16S, <i>cbbL/rbcL</i> and <i>hoxV-hupL</i> library of accretion ice core from 3,561 m | <i>Hydrogenophilus thermoluteolus</i> | Potential hydrogenotrophy | Thermophilic chemolithoautotrophic <i>Hydrogenophilus thermoluteolus</i> 16S rRNA, RuBisCO and NiFe-hydrogenase genes detected. | (Lavire <i>et al.</i> , 2006) |
| Lake Vostok | Largest subglacial lake, isolated from surface 420,000 years | 16S library of Vostok drilling fluid recovered from 4 depths of the bore hole | <i>Sphingomonas</i> , potential contaminants related to human/animal pathogens or saprophytes and environmental contaminants | Hydrocarbon degrading heterotrophs | New contaminant bacteria identified that were associated with hydrocarbon-based drilling fluid. | (Alekhanina <i>et al.</i> , 2007) |

Appendix B

Proteins identified in Ace Lake metaproteomic analysis

Table B.1: Proteins identified in the Ace Lake 5 m sample 0.1 µm size-fraction metaproteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (*a–z, aa–pp*) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

| Gene ID | NSA | COG/NR ID | KO | Locus | 5 m – COG annotated proteins |
|--------------------|---------|-----------|--------|-----------------|---|
| | | | | | COG : KEGG/NR description |
| 167852195 <i>f</i> | 0.02530 | COG1653 | K10232 | AAur_0459 | sugar-binding periplasmic proteins/domains : putative alpha-glucosides-binding ABC transporter (AglE) |
| 167782381* | 0.01724 | COG1879 | K02058 | Ping_2790 | periplasmic sugar-binding proteins : bifunctional carbohydrate binding and transport protein |
| 167813321 | 0.01388 | COG1629 | | GFO_2756 | outer membrane receptor proteins, mostly Fe transport : TonB-dependent outer membrane receptor |
| 167754347 | 0.01044 | COG1879 | K02058 | CMM_0792 | periplasmic sugar-binding proteins : putative sugar ABC transporter, solute-binding protein |
| 167701754 <i>a</i> | 0.00967 | COG0834 | K09969 | SAR11_0953 | ABC-type amino acid transport system, periplasmic component : yhdW |
| 167792775 | 0.00630 | COG1879 | K10552 | SMc02171 | periplasmic sugar-binding proteins : fructose transport system substrate-binding protein |
| 167932252 <i>d</i> | 0.00537 | COG0715 | K02051 | SAR11_0807 | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 167759671 | 0.00493 | COG0605 | | NEMVE_v1g231554 | superoxide dismutase |
| 167751919 <i>h</i> | 0.00468 | COG3740 | | ROP_69760 | phage head maturation protease |
| 167907426 | 0.00438 | COG1638 | | SAR11_0266 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit (mannitol/chloroaromatic compounds) |
| 167711086 | 0.00425 | COG0834 | K02030 | TM1040_0294 | ABC-type amino acid transport system, periplasmic component : lysine-arginine-ornithine-binding periplasmic protein |
| 167819184 | 0.00389 | COG2113 | K02002 | SAR11_1302 | ABC-type proline/glycine betaine transport systems, periplasmic components : opuAC |
| 167680030 | 0.00346 | COG0683 | K01999 | AAur_1271 | ABC-type branched-chain amino acid transport systems, periplasmic component : braC |
| 167865828 <i>b</i> | 0.00338 | COG0834 | K09969 | HCH_05807 | ABC-type amino acid transport system, periplasmic component |
| 167684228 <i>c</i> | 0.00331 | COG0459 | K04077 | SAR11_0162 | chaperonin GroEL (HSP60 family) |
| 167868594 <i>d</i> | 0.00311 | COG0715 | K02051 | SAR11_0807 | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 167785199 <i>c</i> | 0.00309 | COG0459 | K04077 | CHU_1828 | chaperonin GroEL (HSP60 family) |
| 167819050 | 0.00304 | COG2113 | K02001 | Plav_1066 | ABC-type proline/glycine betaine transport systems, periplasmic components |
| 167867034 | 0.00284 | COG0687 | K02055 | SAR11_1336 | spermidine/putrescine-binding periplasmic protein : potD; |
| 167700934 | 0.00277 | COG0450 | | SPO3383 | peroxiredoxin : thiol-specific antioxidant protein |
| 167816084 <i>a</i> | 0.00253 | COG0834 | K09969 | SAR11_0953 | ABC-type amino acid transport system, periplasmic component : yhdW |
| 167714114 | 0.00179 | COG0687 | K02055 | SAR11_1336 | spermidine/putrescine-binding periplasmic protein : potD |
| 167712994 <i>b</i> | 0.00175 | COG0834 | K09969 | HCH_05807 | ABC-type amino acid transport system, periplasmic component |
| 167824568 | 0.00164 | COG3181 | | Dshi_2450 | uncharacterized BCR : hypothetical protein |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------|---------|-----------|--------|---------------|--|
| 167925495 | 0.00159 | COG1653 | K02027 | Krad_1380 | sugar-binding periplasmic proteins/domains |
| 167695410a | 0.00155 | COG0834 | K09969 | SAR11_0953 | ABC-type amino acid transport system, periplasmic component : yhdW |
| 167695984* | 0.00138 | COG1879 | | | periplasmic sugar-binding proteins |
| 167703404 | 0.00134 | COG1012 | K00128 | AAur_pTC20196 | NAD-dependent aldehyde dehydrogenases |
| 167718230 | 0.00125 | COG0683 | K01999 | AAur_1271 | ABC-type branched-chain amino acid transport systems, periplasmic component : braC |
| 167735996 | 0.00103 | COG0591 | | SAR11_0316 | Na+/proline, Na+/panthothenate symporters and related permeases : yjcG |
| 167739054 | 0.00101 | COG1028 | K00059 | SH0230 | dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) : 3-oxoacyl-[acyl-carrier protein] reductase |
| 167701096 | 0.00100 | COG0776 | | KRH_03630 | bacterial nucleoid DNA-binding protein : HU_IHF family transcriptional regulator |
| 167817334 | 0.00098 | COG0715 | | FRAAL1422 | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 167703266c | 0.00095 | COG0459 | K04077 | Noca_3982 | chaperonin GroEL (HSP60 family) |
| 167768609 | 0.00095 | COG3181 | | RD1_2202 | uncharacterized BCR |
| 167868532 | 0.00093 | COG3181 | K07795 | HCH_01639 | uncharacterized BCR : putative tricarboxylic transport membrane protein |
| 167865516 | 0.00088 | COG0747 | | CMM_2185 | ABC-type dipeptide/oligopeptide/nickel transport systems, periplasmic components |
| 167911715 | 0.00083 | COG0776 | K03530 | Sala_0799 | bacterial nucleoid DNA-binding protein : DNA-binding protein HU-beta |
| 167736316 | 0.00082 | COG0174 | K01915 | SAR11_0747 | glutamine synthase : glnA |
| 167916441 | 0.00079 | COG1629 | | BF2044 | outer membrane receptor proteins, mostly Fe transport : putative TonB-dependent outer membrane receptor protein |
| 167920571 | 0.00078 | COG0776 | K03530 | SAR11_0817 | bacterial nucleoid DNA-binding protein : hupA |
| 167703332 | 0.00066 | COG1732 | K05845 | Strop_1633 | periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system (osmoprotectant binding protein) |
| 167662373 | 0.00063 | COG0834 | | Pden_1025 | ABC-type amino acid transport system, periplasmic component : extracellular solute-binding protein, family 3 |
| 167890974 | 0.00062 | COG1878 | | nfa12380 | uncharacterized ACR, predicted metal-dependent hydrolases |
| 167824660 | 0.00061 | COG0683 | K01999 | SAR11_1361 | ABC-type branched-chain amino acid transport systems, periplasmic component : livJ2; Leu/Ile/Val-binding protein precursor |
| 167886240 | 0.00061 | COG0335 | K02884 | CHU_0120 | rplS; 50S ribosomal protein L19 |
| 167921445 | 0.00058 | COG0811 | | GFO_0088 | biopolymer transport proteins : exbB; ExbB-like MotA/TolQ/ExbB family |
| 167776275ee | 0.00055 | COG3740 | K06904 | BL0376 | phage head maturation protease |
| 167659892*ee | 0.00055 | COG3740 | | ROP_69760 | phage head maturation protease |
| 167786475 | 0.00054 | COG0098 | K02988 | Fjoh_0380 | rpsE; 30S ribosomal protein S5 |
| 167693676* | 0.00054 | COG0776 | | Arth_3916 | bacterial nucleoid DNA-binding protein |
| 167818330 | 0.00048 | COG0683 | | Rxyl_0363 | ABC-type branched-chain amino acid transport systems, periplasmic component : extracellular ligand-binding receptor |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|---|
| 167739596 | 0.00044 | COG0545 | K03772 | BDI_2705 | FKBP-type peptidyl-prolyl cis-trans isomerases 1 |
| 167808311c | 0.00044 | COG0459 | K04077 | SAR11_0162 | chaperonin GroEL (HSP60 family) |
| 167706214 | 0.00044 | COG0683 | | Tfu_1779 | ABC-type branched-chain amino acid transport systems, periplasmic component |
| 167866918 | 0.00044 | COG2885 | K03640 | SAR11_0598 | outer membrane protein and related peptidoglycan-associated (lipo)proteins : ompA; OmpA family |
| 167807477 | 0.00042 | COG0834 | | MSMEG_5368 | ABC-type amino acid transport system, periplasmic component : ehuB; ectoine/hydroxyectoine ABC transporter solute-binding protein |
| 167881416e | 0.00040 | COG0050 | K02358 | CHU_3175 | GTPases - translation elongation factors : tufB, tuf |
| 167765645f | 0.00034 | COG1653 | K10232 | Sare_3967 | sugar-binding periplasmic proteins/domains |
| 167730910 | 0.00033 | COG3181 | | Dshi_2450 | uncharacterized BCR |
| 167725574 | 0.00032 | COG0450 | K03386 | CHU_2724 | peroxiredoxin : ahpC; alkyl hydroperoxide reductase, subunit C |
| 167768817 | 0.00032 | COG0740 | K00288 | CHU_1706 | protease subunit of ATP-dependent Clp proteases : methylenetetrahydrofolate dehydrogenase (NADP+) |
| 167886236 | 0.00031 | COG0228 | K02959 | CHU_0117 | rpsP; 30S ribosomal protein S16 |
| 167907528 | 0.00031 | COG0591 | | SAR11_0316 | Na+/proline, Na+/panthothenate symporters and related permeases : yjcG |
| 167868396 | 0.00029 | COG2358 | | PBPRA0389 | predicted periplasmic binding protein : putative immunogenic protein |
| 167718328 | 0.00027 | COG1744 | K07335 | AAur_1253 | surface lipoprotein : basic membrane protein A and related proteins |
| 167769503c | 0.00027 | COG0459 | | CMS_2756 | chaperonin GroEL (HSP60 family) |
| 167818958 | 0.00026 | COG1638 | | TM1040_0356 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit |
| 167702878 | 0.00025 | COG1879 | | Krad_1186 | periplasmic sugar-binding proteins : periplasmic binding protein/LacI transcriptional regulator |
| 167665756 | 0.00025 | COG0091 | K02890 | GFO_2834 | rplV; 50S ribosomal protein L22 |
| 167730894 | 0.00023 | COG1638 | | RD1_2185 | dicarboxylate-binding periplasmic protein : dctP; C4-dicarboxylate-binding periplasmic protein, putative |
| 167680092 | 0.00022 | COG0094 | K02931 | Lxx20210 | rplE; 50S ribosomal protein L5 |
| 167868548 | 0.00020 | COG0834 | K10018 | SAR11_1210 | ABC-type amino acid transport system, periplasmic component : octopine/nopaline transport system substrate-binding protein |
| 167892279 | 0.00019 | COG0834 | K02030 | Veis_2153 | ABC-type amino acid transport system, periplasmic component |
| 167817276 | 0.00018 | COG0347 | K04751 | Acel_1565 | nitrogen regulatory protein PII |
| 167862242 | 0.00018 | COG0087 | K02906 | BT_2727 | rplC; 50S ribosomal protein L3 |
| 167933288 | 0.00018 | COG1638 | | Dshi_3326 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter-DctP subunit |
| 167713980* | 0.00016 | COG0330 | K04088 | SAR11_0008 | membrane protease subunits, stomatin/prohibitin homologs : hflK |
| 167867886 | 0.00016 | COG3181 | | Csal_1767 | uncharacterized BCR : uncharacterized protein UPF0065 |
| 167809873 | 0.00016 | COG0539 | K02945 | CHU_1951 | rpsA; 30S ribosomal protein S1 |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|---------|-----------|--------|------------|---|
| 167713982 | 0.00015 | COG0330 | K04087 | SAR11_0007 | membrane protease subunits, stomatin/prohibitin homologs : hflC |
| 167822210 | 0.00015 | COG0776 | | SCO2950 | bacterial nucleoid DNA-binding protein : hup, SCE59.09c; DNA-binding protein Hu (hs1) |
| 167820450* | 0.00015 | COG1192 | | tlr0963 | ATPases involved in chromosome partitioning : probable cell division inhibitor minD |
| 167820614g | 0.00015 | COG0174 | K01915 | Krad_3291 | glutamine synthase |
| 167714092 | 0.00014 | COG0683 | K01999 | SAR11_1346 | ABC-type branched-chain amino acid transport systems, periplasmic component : livJ |
| 167817130 | 0.00014 | COG0711 | K02109 | SCO5369 | F0F1-type ATP synthase b subunit |
| 167818902 | 0.00014 | COG1638 | | SAR11_0864 | dicarboxylate-binding periplasmic protein |
| 167866078 | 0.00013 | COG0605 | K00518 | Arth_2086 | superoxide dismutase |
| 167865698 | 0.00013 | COG0740 | K01358 | AAur_2381 | protease subunit of ATP-dependent Clp proteases |
| 167817852 | 0.00013 | COG0683 | | Noca_3017 | ABC-type branched-chain amino acid transport systems, periplasmic component : extracellular ligand -binding receptor |
| 167714042 | 0.00012 | COG0683 | K01999 | SAR11_1361 | ABC-type branched-chain amino acid transport systems, periplasmic component : livJ2; Leu/Ile/Val-binding protein precursor |
| 167821000 | 0.00012 | COG1732 | | MSMEG_2924 | periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system (osmoprotectant binding protein) : permease binding-protein component |
| 167668848g | 0.00011 | COG0174 | K01915 | CMM_1636 | glutamine synthase : glnA1 |
| 167748683* | 0.00010 | COG0834 | | PFL_3548 | ABC-type amino acid transport system, periplasmic component |
| 167718146 | 0.00010 | COG0088 | K02926 | Lxx20320 | rplD; 50S ribosomal protein L4 |
| 167862420 | 0.00010 | COG1629 | | FP0112 | outer membrane receptor proteins, mostly Fe transport : probable TonB-dependent outer membrane receptor precursor |
| 167696080* | 0.00010 | COG1344 | K02406 | Csac_1680 | flagellin and related hook-associated proteins |
| 167735512 | 0.00010 | COG0803 | K09815 | Smed_1697 | ABC-type Mn/Zn transport system, periplasmic Mn/Zn-binding (lipo)protein (surface adhesin A) |
| 167719882 | 0.00009 | COG0096 | K02994 | SAR11_1103 | rpsH; 30S ribosomal protein S8 |
| 167660431h | 0.00009 | COG3740 | K06904 | BL0376 | phage head maturation protease |
| 167718250 | 0.00009 | COG2213 | K02799 | GK1948 | phosphotransferase system, mannitol-specific IIBC component |
| 167719862* | 0.00008 | COG0185 | K02965 | SAR11_1113 | rpsS; 30S ribosomal protein S19 |
| 167702806 | 0.00008 | COG0081 | K02863 | KRH_05860 | rplA; 50S ribosomal protein L1 |
| 167719824*e | 0.00008 | COG0050 | K02358 | SAR11_1130 | GTPases - translation elongation factors : tufB, tuf |
| 167817466 | 0.00008 | COG0404 | | mll1258 | glycine cleavage system T protein (aminomethyltransferase) : sarcosine dehydrogenase |
| 167868614 | 0.00008 | COG2113 | K02002 | SAR11_0797 | ABC-type proline/glycine betaine transport systems, periplasmic components : proX |
| 167933120 | 0.00007 | COG0803 | K09815 | Atu1521 | ABC-type Mn/Zn transport system, periplasmic Mn/Zn-binding (lipo)protein (surface adhesin A) : znuA |

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|------------|--|
| 167718168 | 0.00007 | COG0094 | | CMS_0295 | 50S ribosomal protein L5 |
| 167868724 | 0.00007 | COG0396 | K09013 | SAR11_0740 | iron-regulated ABC transporter ATPase subunit SufC |
| 167718156 | 0.00007 | COG0092 | K02982 | Krad_0694 | ribosomal protein S3 |
| 167719956 | 0.00006 | COG0834 | K02030 | SAR11_1068 | ABC-type amino acid transport system, periplasmic component : pheC; cyclohexadienyl dehydratase |
| 167730882 | 0.00006 | COG0004 | | SAR11_0818 | ammonia permeases : amtB; ammonium transporter |
| 167718138e | 0.00006 | COG0050 | K02358 | Tfu_2648 | GTPases - translation elongation factors: tuf |
| 167718052 | 0.00006 | COG1653 | K02027 | Krad_3469 | sugar-binding periplasmic proteins/domains |
| 167700960* | 0.00006 | COG3794 | | SMa1243 | plastocyanin : Azu1 pseudoazurin (blue copper protein) |
| 167868482 | 0.00005 | COG0715 | | | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 167868656 | 0.00005 | COG0715 | K02051 | AZC_2351 | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 167868494 | 0.00004 | COG1638 | | SMa0157 | dicarboxylate-binding periplasmic protein |
| 167866460 | 0.00004 | COG0687 | K02055 | SCO5667 | spermidine/putrescine-binding periplasmic protein |
| 167719840 | 0.00004 | COG0085 | K03043 | SAR11_1123 | DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mtthe, Aful) : rpoB |
| 167816636 | 0.00004 | COG0459 | K04077 | Krad_0736 | chaperonin GroEL (HSP60 family) |
| 167701680* | 0.00004 | COG3740 | K06904 | BL0376 | phage head maturation protease |
| 167717794* | 0.00003 | COG0195 | K02600 | SAR11_0388 | phage head maturation protease |
| 167717838 | 0.00003 | COG0443 | K04043 | SAR11_0368 | molecular chaperone : dnaK |
| 167834314 | 0.00003 | COG0443 | | CMS_2806 | molecular chaperone : dnaK |
| 167717784 | 0.00003 | COG1185 | K00962 | SAR11_0392 | polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) |
| 167818278 | 0.00003 | COG1022 | | NoCa_3113 | long-chain acyl-CoA synthetases (AMP-forming) : AMP-dependent synthetase and ligase |
| 167719850 | 0.00002 | COG0480 | K02355 | SAR11_1119 | translation elongation and release factors (GTPases) : fusA |
| 167816480 | 0.00002 | COG0086 | K03046 | Krad_0681 | DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn) |
| 167866408 | 0.00001 | COG1185 | K00962 | Lxx09030 | polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) |

5 m – KEGG and NR annotated proteins

| | | | | |
|--------------------|---------|--------------|-----------|---|
| 167873078 | 0.03710 | BAF91544 | | major capsid protein [uncultured Myoviridae] |
| 167771989 <i>i</i> | 0.02658 | | BTH_I0914 | hypothetical protein |
| 167723550 <i>j</i> | 0.01559 | YP_001648266 | | hypothetical protein [Ostreococcus virus OsV5] |
| 167927818 <i>j</i> | 0.01345 | YP_001648266 | | hypothetical protein [Ostreococcus virus OsV5] |
| 167933090 | 0.01298 | YP_002590925 | | putative porin [Candidatus Pelagibacter sp. HTCC7211] |
| 167711088 | 0.01230 | | K09969 | putative amino acid ABC transporter, periplasmic amino acid-binding protein |
| 167691398 | 0.01175 | | HMI_2880 | phage major capsid protein, hk97 family |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---------------------|---------|--------------|----|--------------|--|
| 167922719 <i>k</i> | 0.01174 | | | Neut_1469 | phage major capsid protein, HK97 family protein |
| 167775105 <i>l</i> | 0.01103 | ABC95191 | | | GP23-major capsid protein [Stenotrophomonas phage SMB14] |
| 167687982 <i>l</i> | 0.00968 | NP_944113 | | | gp23 major head protein [Aeromonas phage Aeh1] |
| 167748599 <i>m</i> | 0.00960 | | | M6_Spy1138 | phage prohead protease |
| 167733772 | 0.00952 | | | BBta_5785 | putative phage major head protein |
| 167925660 | 0.00923 | | | SRU_2178 | putative outer membrane protein, probably involved in nutrient binding |
| 167796059 <i>n</i> | 0.00853 | | | APECO1_525 | hypothetical protein |
| 167883590 | 0.00820 | | | PP_1567 | phage major capsid protein, HK97 family |
| 167666520 <i>o</i> | 0.00818 | BAF91544 | | | major capsid protein [uncultured Myoviridae] |
| 167664173* <i>p</i> | 0.00713 | | | GDI3673 | hypothetical protein |
| 167667150 <i>m</i> | 0.00687 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 167771337 | 0.00605 | ABW90951 | | | gp23 major capsid protein [uncultured Myoviridae] |
| 167884290 <i>l</i> | 0.00573 | BAF91544 | | | major capsid protein [uncultured Myoviridae] |
| 167760139 | 0.00561 | | | CHU_2679 | probable outer membrane lipoprotein P61 |
| 167816468 <i>q</i> | 0.00522 | | | DR_A0099 | hypothetical protein |
| 167700634 <i>j</i> | 0.00499 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167687792 <i>r</i> | 0.00498 | | | Asuc_1240 | phage major capsid protein, HK97 family |
| 167842580 <i>r</i> | 0.00453 | | | BAV1464 | major capsid protein |
| 167700776 | 0.00447 | | | Bpro_3745 | hypothetical protein |
| 167729766 | 0.00433 | ZP_01224596 | | GB2207_03424 | hypothetical protein [marine gamma proteobacterium HTCC2207] |
| 167934698 | 0.00431 | | | Swit_4452 | hypothetical protein |
| 167884738 | 0.00409 | | | BBta_5785 | putative phage major head protein |
| 167669610 <i>p</i> | 0.00397 | | | GDI3673 | hypothetical protein |
| 167861688 | 0.00359 | | | BDI_2874 | putative outer membrane protein, probably involved in nutrient binding |
| 167910063 | 0.00347 | | | GDI3673 | hypothetical protein |
| 167893743* <i>s</i> | 0.00338 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167888926 <i>p</i> | 0.00326 | | | GDI3673 | hypothetical protein |
| 167753643* | 0.00324 | | | Daci_1946 | putative phage major head protein |
| 167742624 <i>p</i> | 0.00317 | | | GDI3673 | hypothetical protein |
| 167908539 <i>r</i> | 0.00304 | | | BAV1464 | major capsid protein |
| 167675286* | 0.00284 | | | CKO_01864 | hypothetical protein |
| 167900893 <i>n</i> | 0.00278 | | | APECO1_525 | hypothetical protein |
| 167778265 <i>p</i> | 0.00275 | | | GDI3673 | hypothetical protein |

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|---------|--------------|----|-------------|---|
| 167786471 | 0.00267 | | | mlr8524 | phage major capsid protein, GP36 |
| 167735768 | 0.00265 | | | FRAAL2681 | hypothetical protein |
| 167773951t | 0.00253 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 167841586j | 0.00250 | A7U6E7 | | | putative major capsid protein [Chrysochromulina ericina virus] |
| 167919545 | 0.00245 | | | Pmen_3970 | phage major capsid protein, HK97 family |
| 167781901u | 0.00236 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 167923659p | 0.00235 | | | GDI3673 | hypothetical protein |
| 167852301v | 0.00230 | | | MAB_1788 | bacteriophage protein |
| 167659301 | 0.00224 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 167861686 | 0.00223 | YP_002789013 | | | TonB dependent/ligand-gated channel [Polaribacter sp. MED152] |
| 167712528v | 0.00215 | | | MAB_1788 | bacterial nucleoid DNA-binding protein |
| 167678920*w | 0.00209 | YP_001498525 | | AR158_C444L | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 167849540j | 0.00201 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167781903u | 0.00201 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 167863158j | 0.00200 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167663967* | 0.00190 | | | Swit_4461 | hypothetical protein |
| 167687108u | 0.00182 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 167692622i | 0.00176 | | | SG1188 | hypothetical protein |
| 167733858 | 0.00170 | ZP_01017474 | | | major capsid protein, HK97 family protein [Parvularcula bermudensis HTCC2503] |
| 167852851p | 0.00167 | | | GDI3673 | hypothetical protein |
| 167864542k | 0.00166 | | | Neut_1469 | phage major capsid protein, HK97 family protein |
| 167803157 | 0.00165 | ZP_01688540 | | | lipoprotein, putative [Microscilla marina ATCC 23134] |
| 167682644j | 0.00153 | A7U6E7 | | | putative major capsid protein [Chrysochromulina ericina virus] |
| 167733004*j | 0.00150 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 167765429 | 0.00148 | | | CHU_2610 | gliding motility-related protein; possible GldN and/or GldO |
| 167878228t | 0.00145 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 167775103 | 0.00145 | YP_214669 | | | gp23 [Prochlorococcus phage P-SSM4] |
| 167702908 | 0.00143 | | | mma_2202 | hypothetical protein |
| 167869946 | 0.00135 | YP_195142 | | | major capsid protein gp23 [Synechococcus phage S-PM2] |
| 167834518 | 0.00128 | | | Haur_0657 | hypothetical protein |
| 167807747 | 0.00122 | | | Saro_0657 | hypothetical protein |
| 167816420 | 0.00121 | | | APECO1_525 | hypothetical protein |

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------|---------|--------------|--------|----------------|---|
| 167809283k | 0.00119 | | | Neut_1469 | phage major capsid protein, HK97 family protein |
| 167868514 | 0.00119 | | | SAR11_1290 | TRAP-type bacterial extracellular solute-binding protein |
| 167750765 | 0.00118 | | | Smed_1334 | phage major capsid protein, HK97 family |
| 167925457p | 0.00115 | | | GDI3673 | hypothetical protein |
| 167782759 | 0.00113 | | | Oter_1957 | band 7 protein |
| 167871794j | 0.00113 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167756019 | 0.00112 | | | CHU_0172 | gldL; gliding motility-related protein |
| 167670926x | 0.00112 | | | BBta_5785 | putative phage major head protein |
| 167821362z | 0.00112 | | | APECO1_525 | hypothetical protein |
| 167690910j | 0.00111 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167685332j | 0.00110 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167700460* | 0.00110 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167685474*aa | 0.00104 | YP_001648182 | | OsV5_105r | hypothetical protein [Ostreococcus virus OsV5] |
| 167734326bb | 0.00103 | | | amb4267 | hypothetical protein |
| 16775653q | 0.00101 | | | Haur_0657 | hypothetical protein |
| 167733302p | 0.00101 | | | GDI3673 | hypothetical protein |
| 167663981p | 0.00097 | | | GDI3673 | hypothetical protein |
| 167763843 | 0.00096 | | | Saro_0657 | hypothetical protein |
| 167768193z | 0.00096 | | | CKO_01864 | hypothetical protein |
| 167719228i | 0.00095 | | | SG1188 | hypothetical protein |
| 167844676 | 0.00091 | ZP_03643684 | | BACCOPRO_02057 | hypothetical protein [Bacteroides coprophilus DSM 18228] |
| 167881504cc | 0.00091 | | | BSU26140 | yqbE; hypothetical protein |
| 167852559oo | 0.00090 | | | HSM_0907 | hypothetical protein |
| 167804465* | 0.00088 | ZP_03724502 | | ObacDRAFT_9001 | hypothetical protein [Opitutaceae bacterium TAV2] |
| 167794165p | 0.00087 | | | GDI3673 | hypothetical protein |
| 167734676 | 0.00085 | | | Amet_4028 | phage major capsid protein, HK97 family |
| 167764813u | 0.00084 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 167781783p | 0.00079 | | | GDI3673 | hypothetical protein |
| 167759955 | 0.00077 | | | Dgeo_0628 | hypothetical protein |
| 167733674 | 0.00076 | | | Swit_4452 | hypothetical protein |
| 167878828 | 0.00075 | | K02027 | SAV1394 | ABC transporter solute-binding protein |
| 167740142 | 0.00075 | YP_002705257 | | | gp34 [Stenotrophomonas sp. SKA14] |
| 167834088 | 0.00074 | | | Haur_0657 | hypothetical protein |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|---------|--------------|--------|----------------|---|
| 167821604j | 0.00072 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167798697p | 0.00070 | | | GDI3673 | hypothetical protein |
| 167823322s | 0.00070 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167718758j | 0.00070 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 167713806 | 0.00069 | | | SG1188 | hypothetical protein |
| 167778269dd | 0.00068 | YP_002276820 | | Gdia_2460 | hypothetical protein [Gluconacetobacter diazotrophicus PA1 5] |
| 167879936s | 0.00067 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167701850 | 0.00064 | | | M446_5960 | hypothetical protein |
| 167934792p | 0.00064 | | | GDI3673 | hypothetical protein |
| 167874674 | 0.00063 | | | GFO_0492 | conserved hypothetical protein, secreted-possibly porin |
| 167821292 | 0.00062 | | | Oant_1504 | peptidase U35 phage prohead HK97 |
| 167867556 | 0.00062 | | | Rru_A2587 | hypothetical protein |
| 167901481 | 0.00061 | | | Cthe_1719 | phage major capsid protein, HK97 family |
| 167824444 | 0.00059 | | | Smed_5134 | TRAP dicarboxylate transporter-DctP subunit |
| 167696166* | 0.00059 | | | BTH_I0915 | hypothetical protein |
| 167936648 | 0.00056 | EEI06235 | | XcelDRAFT_1815 | hypothetical protein [Xylanimonas cellulosilytica DSM 15894] |
| 167910361j | 0.00054 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 167675492*p | 0.00054 | | | GDI3673 | hypothetical protein |
| 167801933aa | 0.00052 | YP_001648182 | | OsV5_105r | hypothetical protein [Ostreococcus virus OsV5] |
| 167725772cc | 0.00051 | | | BSU26140 | yqbE; hypothetical protein |
| 167820670 | 0.00050 | | | gll0198 | similar to bacteriorhodopsin |
| 167832972t | 0.00050 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 167867536 | 0.00049 | | | TM1040_0812 | hypothetical protein |
| 167783747 | 0.00048 | | | Glov_2914 | cell surface receptor IPT/TIG domain protein |
| 167893945 | 0.00048 | | | Oter_3420 | hypothetical protein |
| 167740708 | 0.00047 | | K03286 | Pnap_1319 | OmpA/MotB domain protein; OmpA-OmpF porin, OOP family |
| 167772783 | 0.00047 | | | RCIX1696 | hypothetical protein |
| 167734614p | 0.00047 | | | GDI3673 | hypothetical protein |
| 167776587* | 0.00046 | YP_001648249 | | OsV5_172f | hypothetical protein [Ostreococcus virus OsV5] |
| 167911245s | 0.00046 | YP_001648315 | | OsV5_239r | hypothetical protein [Ostreococcus virus OsV5] |
| 167867748x | 0.00044 | | | Daci_1946 | putative phage major head protein |
| 167732694 | 0.00044 | | | NMC0858 | putative phage-related protein |
| 167922873 | 0.00042 | | | CHU_3230 | hypothetical protein |

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------|---------|--------------|--------|-----------------|--|
| 167734178p | 0.00041 | | | GDI3673 | hypothetical protein |
| 167873260 | 0.00041 | | | Sare_3763 | hypothetical protein |
| 167685638*w | 0.00041 | YP_001498525 | | AR158_C444L | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 167901149 | 0.00040 | | K01358 | azo1870 | endopeptidase Clp; K01358 ATP-dependent Clp protease, protease subunit |
| 167853099i | 0.00040 | | | SG1188 | hypothetical protein |
| 167761349x | 0.00037 | | | Daci_1946 | putative phage major head protein |
| 167776241 | 0.00037 | | | mll0455 | hypothetical protein |
| 167824154 | 0.00036 | | | SACE_4894 | hydrolase, alpha/beta fold family |
| 167843578s | 0.00035 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167682808* | 0.00035 | YP_001648239 | | OsV5_162f | hypothetical protein [Ostreococcus virus OsV5] |
| 167703228 | 0.00034 | | | Dshi_0412 | beta-Ig-H3/fasciclin |
| 167918033p | 0.00034 | | | GDI3673 | hypothetical protein |
| 167891224p | 0.00033 | | | GDI3673 | hypothetical protein |
| 167867622 | 0.00033 | | | Oant_1504 | peptidase U35 phage prohead HK97 |
| 167732430ff | 0.00031 | ZP_02092868 | | FAEPRf212_03171 | hypothetical protein [Faecalibacterium prausnitzii M21/2] |
| 167684500s | 0.00031 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167824164* | 0.00030 | YP_001648240 | | OsV5_163f | hypothetical protein [Ostreococcus virus OsV5] |
| 167765431 | 0.00030 | | | CHU_0173 | gldM; gliding motility-related protein |
| 167854137 | 0.00030 | ZP_00743477 | | RBTH_08297 | hypothetical protein [Bacillus thuringiensis serovar israelensis ATCC 35646] |
| 167730288k | 0.00028 | | | Neut_1469 | phage major capsid protein, HK97 family protein |
| 167685472*gg | 0.00027 | YP_001648184 | | OsV5_107r | hypothetical protein [Ostreococcus virus OsV5] |
| 167778267p | 0.00027 | | | GDI3673 | hypothetical protein |
| 167919557 | 0.00027 | | | PputW619_3936 | hypothetical protein |
| 167782645 | 0.00026 | | | BBta_5785 | putative phage major head protein |
| 167908551 | 0.00026 | | | CC_2781 | hypothetical protein |
| 16786723p | 0.00026 | | | GDI3673 | hypothetical protein |
| 167896531 | 0.00026 | | | BAV1464 | major capsid protein |
| 167821374 | 0.00025 | YP_001919460 | | Mpop_5468 | hypothetical protein [Methylobacterium populi BJ001] |
| 167833472 | 0.00024 | | | GDI3673 | hypothetical protein |
| 167733210 | 0.00024 | | | Pmen_3970 | phage major capsid protein, HK97 family |
| 167713652* | 0.00023 | | | mlr8533 | hypothetical protein |
| 167935700p | 0.00023 | | | GDI3673 | hypothetical protein |
| 167872214hh | 0.00023 | | K06907 | Sfum_3815 | phage tail sheath protein |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------|---------|--------------|--------|------------------|---|
| 167881636 | 0.00023 | | | Rspf17025_0103 | hypothetical protein |
| 167791200p | 0.00023 | | | GDI3673 | hypothetical protein |
| 167922981p | 0.00022 | | | GDI3673 | hypothetical protein |
| 167824604 | 0.00022 | | | Rspf17029_3578 | uncharacterized protein UPF0065 |
| 167933608ff | 0.00022 | ZP_02092868 | | FAEPRAM212_03171 | hypothetical protein [Faecalibacterium prausnitzii M21/2] |
| 167817058 | 0.00022 | | K00518 | Sare_4077 | superoxide dismutase |
| 167823358*gg | 0.00021 | YP_001648184 | | OsV5_107r | hypothetical protein [Ostreococcus virus OsV5] |
| 167892855 | 0.00021 | YP_001648184 | | OsV5_107r | hypothetical protein [Ostreococcus virus OsV5] |
| 167840790 | 0.00021 | | | Rspf17025_0437 | hypothetical protein |
| 167712150*s | 0.00021 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167833160bb | 0.00021 | | | amb4267 | hypothetical protein |
| 167892985j | 0.00021 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167696294oo | 0.00020 | | | HS_1377 | hypothetical protein |
| 167759041 | 0.00019 | | | PP_3877 | hypothetical protein |
| 167766087s | 0.00019 | YP_001648153 | | OsV5_076f | hypothetical protein [Ostreococcus virus OsV5] |
| 167919775 | 0.00019 | YP_001294637 | | ORF044 | hypothetical protein [Pseudomonas phage PA11] |
| 167804453* | 0.00017 | ZP_03724505 | | ObacDRAFT_9004 | hypothetical protein [Opitutaceae bacterium TAV2] |
| 167833104 | 0.00017 | | | Bcep1808_1173 | hypothetical protein |
| 167721370* | 0.00016 | YP_001648301 | | OsV5_225r | hypothetical protein [Ostreococcus virus OsV5] |
| 167826943* | 0.00016 | | | Sare_3763 | hypothetical protein |
| 167865492 | 0.00015 | | K02027 | Pput_3473 | extracellular solute-binding protein, family 1; multiple sugar transport system substrate-binding protein |
| 167685780*j | 0.00015 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 167910713 | 0.00014 | | | Bd1641 | hypothetical protein |
| 167910061p | 0.00014 | | | GDI3673 | hypothetical protein |
| 167833358s | 0.00013 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167692314*w | 0.00012 | YP_001498525 | | AR158_C444L | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 167925393 | 0.00012 | | | Oter_3421 | hypothetical protein |
| 167687436 | 0.00011 | | K01999 | azo3443 | conserved hypothetical ABC-type branched-chain amino acid transport systems, periplasmic component |
| 167719658hh | 0.00011 | | K06907 | Dde_1889 | hypothetical protein |
| 167668360s | 0.00011 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167735772 | 0.00010 | | | FRAAL2683 | hypothetical protein; putative mycobacteriophage protein (GP15) similarity |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|---------|--------------|--------|--------------|---|
| 167702102 | 0.00010 | | | Daci_1946 | putative phage major head protein |
| 167688622p | 0.00009 | | | GDI3673 | hypothetical protein |
| 167782867cc | 0.00009 | | | BSU26140 | yqbE; hypothetical protein |
| 167867386 | 0.00008 | | | TM1040_1299 | peptidase U35, phage prohead HK97 |
| 167789595 | 0.00008 | | | APEC01_4044 | hypothetical protein |
| 167828425*w | 0.00007 | YP_001498525 | | AR158_C444L | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 167865490 | 0.00007 | | K02027 | Rmet_2229 | extracellular solute-binding protein, family 1; multiple sugar transport system substrate-binding protein |
| 167840812 | 0.00006 | ZP_01959135 | | BACCAC_00731 | hypothetical protein [Bacteroides caccae ATCC 43185] |
| 167706428 | 0.00005 | YP_001648190 | | OsV5_113r | hypothetical protein [Ostreococcus virus OsV5] |
| 167867920p | 0.00005 | | | GDI3673 | hypothetical protein |
| 167842648*s | 0.00005 | YP_001648315 | | OsV5_239r | hypothetical protein [Ostreococcus virus OsV5] |
| 167869096w | 0.00005 | YP_001498525 | | AR158_C444L | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 167859444 | 0.00005 | YP_001648151 | | OsV5_074f | hypothetical protein [Ostreococcus virus OsV5] |
| 167871600 | 0.00005 | YP_001648152 | | OsV5_075f | hypothetical protein [Ostreococcus virus OsV5] |
| 167669608p | 0.00004 | | | GDI3673 | hypothetical protein |
| 167678686p | 0.00004 | | | GDI3673 | hypothetical protein |
| 167752119 | 0.00004 | YP_001648152 | | OsV5_075f | hypothetical protein [Ostreococcus virus OsV5] |
| 167825992j | 0.00003 | A7U6E7 | | | putative major capsid protein [Chrysochromulina ericina virus |
| 167818634 | 0.00003 | | | PputGB1_1751 | hypothetical protein |
| 167671778* | 0.00003 | YP_001648124 | | OsV5_047f | hypothetical protein [Ostreococcus virus OsV5] |
| 167871626s | 0.00002 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 167753841 | 0.00002 | | K06907 | Sfum_3815 | phage tail sheath protein |
| 167724632* | 0.00002 | YP_001648232 | | OsV5_155f | hypothetical protein [Ostreococcus virus OsV5] |
| 167690816 | 0.00002 | YP_001648190 | | OsV5_113r | hypothetical protein [Ostreococcus virus OsV5] |
| 167742884* | 0.00001 | | | Dvul_0646 | hypothetical protein |
| 167875342j | 0.00001 | YP_001648145 | | OsV5_068f | hypothetical protein [Ostreococcus virus OsV5] |

5 m – Proteins with no annotation

| | |
|-------------|---------|
| 167699580* | 0.01263 |
| 167736790ii | 0.01043 |
| 167796769 | 0.00914 |
| 167722626jj | 0.00789 |
| 167703824pp | 0.00714 |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------------|---------|-----------|----|-------|---------------------------|
| 167753801 | 0.00626 | | | | |
| 167854251 | 0.00577 | | | | |
| 167744898* <i>o</i> | 0.00546 | | | | |
| 167664175* <i>nn</i> | 0.00514 | | | | |
| 167829145 <i>pp</i> | 0.00433 | | | | |
| 167779175 <i>pp</i> | 0.00423 | | | | |
| 167881060 <i>pp</i> | 0.00419 | | | | |
| 167887022 <i>pp</i> | 0.00390 | | | | |
| 167836216 <i>jj</i> | 0.00369 | | | | |
| 167688044 <i>p</i> | 0.00353 | | | | |
| 167697984 <i>pp</i> | 0.00321 | | | | |
| 167855765 <i>jj</i> | 0.00318 | | | | |
| 167764897 <i>jj</i> | 0.00297 | | | | |
| 167718436 | 0.00240 | | | | |
| 167771817 | 0.00226 | | | | |
| 167699330* | 0.00220 | | | | |
| 167891152 <i>pp</i> | 0.00207 | | | | |
| 167844558 <i>pp</i> | 0.00197 | | | | |
| 167801097 <i>pp</i> | 0.00197 | | | | |
| 167891908 | 0.00196 | | | | |
| 167820168 <i>ii</i> | 0.00192 | | | | |
| 167688624 | 0.00182 | | | | |
| 167746546 <i>jj</i> | 0.00176 | | | | |
| 167682238 <i>p</i> | 0.00175 | | | | |
| 167722606 | 0.00164 | | | | |
| 167883488 <i>pp</i> | 0.00157 | | | | |
| 167839862 <i>mm</i> | 0.00139 | | | | |
| 167820406* | 0.00139 | | | | |
| 167858104 | 0.00138 | | | | |
| 167806741 <i>jj</i> | 0.00138 | | | | |
| 167678192 | 0.00133 | | | | |
| 167706644 | 0.00133 | | | | |
| 167787801* <i>ll</i> | 0.00124 | | | | |

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|---------|-----------|----|-------|---------------------------|
| 167781039 | 0.00124 | | | | |
| 167936638 | 0.00116 | | | | |
| 167733554 | 0.00116 | | | | |
| 167918031 | 0.00112 | | | | |
| 167790652 | 0.00110 | | | | |
| 167734428 | 0.00102 | | | | |
| 167925455 | 0.00102 | | | | |
| 167928078 | 0.00100 | | | | |
| 167682970o | 0.00098 | | | | |
| 167701282 | 0.00091 | | | | |
| 167867140o | 0.00087 | | | | |
| 167809975jj | 0.00087 | | | | |
| 167750727 | 0.00082 | | | | |
| 16783564ll | 0.00082 | | | | |
| 167789467pp | 0.00080 | | | | |
| 167669606 | 0.00078 | | | | |
| 167733300 | 0.00076 | | | | |
| 167750389jj | 0.00072 | | | | |
| 167852849 | 0.00072 | | | | |
| 167827017 | 0.00070 | | | | |
| 167691436 | 0.00068 | | | | |
| 167816466* | 0.00067 | | | | |
| 167678688 | 0.00063 | | | | |
| 167796679 | 0.00062 | | | | |
| 167761163 | 0.00059 | | | | |
| 167916021pp | 0.00059 | | | | |
| 167867918 | 0.00058 | | | | |
| 167853885 | 0.00058 | | | | |
| 167757667 | 0.00053 | | | | |
| 167922983 | 0.00052 | | | | |
| 167923663nn | 0.00051 | | | | |
| 167922109 | 0.00051 | | | | |
| 167661777 | 0.00051 | | | | |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---------------------|---------|-----------|----|-------|---------------------------|
| 167936684 | 0.00050 | | | | |
| 167867228 | 0.00050 | | | | |
| 167791202 | 0.00049 | | | | |
| 167819274 | 0.00046 | | | | |
| 167765833* | 0.00045 | | | | |
| 167793451* | 0.00045 | | | | |
| 167732910 | 0.00043 | | | | |
| 167890226 | 0.00043 | | | | |
| 167718438 | 0.00042 | | | | |
| 167688708 | 0.00041 | | | | |
| 167699600 <i>pp</i> | 0.00041 | | | | |
| 167746630 | 0.00040 | | | | |
| 167821290 | 0.00039 | | | | |
| 167916161* | 0.00038 | | | | |
| 167700778 | 0.00037 | | | | |
| 167701632 <i>kk</i> | 0.00037 | | | | |
| 167675494* | 0.00036 | | | | |
| 167711820 | 0.00034 | | | | |
| 167663983 | 0.00033 | | | | |
| 167689444* | 0.00032 | | | | |
| 167933464 <i>nn</i> | 0.00032 | | | | |
| 167891222 | 0.00032 | | | | |
| 167852557 | 0.00031 | | | | |
| 167843828 | 0.00031 | | | | |
| 167843020 | 0.00031 | | | | |
| 167677672 | 0.00030 | | | | |
| 167776503* | 0.00028 | | | | |
| 167804815 | 0.00027 | | | | |
| 167713808 | 0.00027 | | | | |
| 167702310 <i>dd</i> | 0.00026 | | | | |
| 167913463o | 0.00025 | | | | |
| 167881302 | 0.00025 | | | | |
| 167907624 <i>mm</i> | 0.00025 | | | | |

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---------------------|---------|-----------|----|-------|---------------------------|
| 167753609* | 0.00024 | | | | |
| 167829571 | 0.00024 | | | | |
| 167921665 | 0.00024 | | | | |
| 167920645 | 0.00023 | | | | |
| 167714058 | 0.00022 | | | | |
| 167677546* | 0.00021 | | | | |
| 167913465 | 0.00020 | | | | |
| 167697624* | 0.00020 | | | | |
| 167912083 <i>jj</i> | 0.00019 | | | | |
| 167766043* | 0.00018 | | | | |
| 167678558* | 0.00018 | | | | |
| 167733556 | 0.00017 | | | | |
| 167663383* | 0.00016 | | | | |
| 167905220* | 0.00015 | | | | |
| 167891594 | 0.00015 | | | | |
| 167883594 <i>ll</i> | 0.00015 | | | | |
| 167879460 | 0.00014 | | | | |
| 167919777 | 0.00014 | | | | |
| 167884588 <i>o</i> | 0.00014 | | | | |
| 167822810 | 0.00013 | | | | |
| 167713494 | 0.00012 | | | | |
| 167841896* | 0.00010 | | | | |
| 167804467* | 0.00010 | | | | |
| 167925043 <i>kk</i> | 0.00010 | | | | |
| 167858106 | 0.00009 | | | | |
| 167788223* | 0.00009 | | | | |
| 167878206 | 0.00008 | | | | |
| 167764895 <i>jj</i> | 0.00008 | | | | |
| 167767179 | 0.00008 | | | | |
| 167858640 | 0.00008 | | | | |
| 167683530* | 0.00007 | | | | |
| 167918035 | 0.00006 | | | | |
| 167766125* | 0.00006 | | | | |

Continued on next page

Table B.1 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|----|-------|---------------------------|
| 167752051* | 0.00005 | | | | |
| 167890228 | 0.00004 | | | | |
| 167685654* | 0.00003 | | | | |
| 167719670* | 0.00002 | | | | |
| 167879450 | 0.00002 | | | | |

Table B.2: Proteins identified in the Ace Lake 11.5 m sample 0.1 μ m size-fraction metaproteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (a-z, aa-pp) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

| Gene ID | NSA | COG/NR ID | KO | Locus | 11.5 m – COG annotated proteins |
|------------|---------|-----------|--------|-------------|---|
| | | | | | COG : KEGG/NR description |
| 163207432 | 0.01734 | COG0834 | K09969 | SAR11_0953 | ABC-type amino acid transport system, periplasmic component : yhdW |
| 163201696 | 0.01011 | COG1879 | | MSMEG_1374 | periplasmic sugar-binding proteins : ribose ABC transporter, periplasmic binding protein |
| 163136433 | 0.00671 | COG3409 | | Clos_2845 | putative peptidoglycan-binding domain-containing protein |
| 163539247 | 0.00629 | COG1653 | | Noca_3914 | sugar-binding periplasmic proteins/domains : extracellular solute-binding protein, family 1 |
| 163377029a | 0.00566 | COG0050 | K02358 | amb3148 | GTPases - translation elongation factors : tuf |
| 163451248b | 0.00541 | COG0715 | K02051 | SAR11_0807 | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 163135049c | 0.00533 | COG1638 | | SAR11_0266 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit (mannitol/chloroaromatic compounds) |
| 163451084 | 0.00531 | COG2113 | K02002 | SAR11_1302 | ABC-type proline/glycine betaine transport systems, periplasmic components : opuAC |
| 163117735 | 0.00526 | COG2113 | K02001 | Plav_1066 | ABC-type proline/glycine betaine transport systems, periplasmic components |
| 163198494d | 0.00415 | COG0591 | | SAR11_0316 | Na+/proline, Na+/panthothenate symporters and related permeases : yjcG |
| 163416423* | 0.00371 | COG0776 | K03530 | SAR11_0817 | bacterial nucleoid DNA-binding protein : hupA |
| 163442042 | 0.00364 | COG0687 | K02055 | SAR11_1336 | spermidine/putrescine-binding periplasmic protein : potD |
| 163208342e | 0.00356 | COG0834 | K09969 | HCH_05807 | ABC-type amino acid transport system, periplasmic component |
| 163234668* | 0.00338 | COG0450 | | SPO3383 | peroxiredoxin : thiol-specific antioxidant protein |
| 163261506 | 0.00336 | COG1638 | | SAR11_0864 | dicarboxylate-binding periplasmic protein |
| 163104605 | 0.00336 | COG2213 | K02799 | GK1948 | phosphotransferase system, mannitol-specific IIBC component |
| 163357996c | 0.00321 | COG1638 | | SAR11_0266 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit (mannitol/chloroaromatic compounds) |
| 163381848f | 0.00314 | COG0459 | K04077 | SAR11_0162 | chaperonin GroEL (HSP60 family) |
| 163145053 | 0.00312 | COG0683 | K01999 | SAR11_1361 | ABC-type branched-chain amino acid transport systems, periplasmic component : livJ2; Leu/Ile/Val-binding protein precursor |
| 163388714 | 0.00282 | COG1638 | | RD1_2185 | dicarboxylate-binding periplasmic protein : DctP; C4-dicarboxylate-binding periplasmic protein, putative |
| 163450920 | 0.00247 | COG0683 | K01999 | SAR11_1361 | ABC-type branched-chain amino acid transport systems, periplasmic component : livJ2; Leu/Ile/Val-binding protein precursor |
| 163240441 | 0.00246 | COG1653 | K02027 | PflO1_3630 | sugar-binding periplasmic proteins/domains |
| 163275955 | 0.00241 | COG1638 | | TM1040_0356 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter - DctP subunit |

Continued on next page

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---------------------|---------|-----------|--------|----------------|---|
| 163449626 | 0.00229 | COG1638 | | Dshi_3326 | dicarboxylate-binding periplasmic protein : TRAP dicarboxylate transporter, DctP subunit |
| 163450966 | 0.00228 | COG0687 | K02055 | SAR11_1336 | spermidine/putrescine-binding periplasmic protein : potD |
| 163174786 | 0.00223 | COG2358 | | PBPRA0389 | predicted periplasmic binding protein : putative immunogenic protein |
| 163120641 | 0.00219 | COG1879 | K02058 | CMM_0792 | periplasmic sugar-binding proteins : putative sugar ABC transporter, solute-binding protein |
| 163416343 | 0.00204 | COG3181 | | Csal_1767 | uncharacterized BCR |
| 163441934 | 0.00198 | COG2885 | K03640 | SAR11_0598 | outer membrane protein and related peptidoglycan-associated (lipo)proteins : ompA; OmpA family |
| 163320067 | 0.00176 | COG2165 | K02650 | SAR11_0054 | general secretory pathway proteins G and H and related periplasmic/secreted proteins : pilA; pilin (bacterial filament) |
| 163274197 <i>b</i> | 0.00167 | COG0715 | K02051 | SAR11_0807 | ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, periplasmic components |
| 163128105 | 0.00167 | COG1012 | K00128 | AAur_pTC20196 | NAD-dependent aldehyde dehydrogenases |
| 163214443 <i>e</i> | 0.00161 | COG0834 | K09969 | HCH_05807 | ABC-type amino acid transport system, periplasmic component |
| 163174178 | 0.00143 | COG1879 | | RHA1_ro08504 | periplasmic sugar-binding proteins : ABC sugar transporter, periplasmic substrate binding protein |
| 163134937 <i>d</i> | 0.00138 | COG0591 | | SAR11_0316 | Na+/proline, Na+/panthothenate symporters and related permeases : yjcG |
| 163451228 | 0.00132 | COG2113 | K02002 | SAR11_0797 | ABC-type proline/glycine betaine transport systems, periplasmic components : proX |
| 163376697* | 0.00129 | COG0834 | K10018 | SAR11_1210 | ABC-type amino acid transport system, periplasmic component : octopine/nopaline transport system substrate-binding protein |
| 163104625 | 0.00119 | COG0683 | K01999 | AAur_1271 | ABC-type branched-chain amino acid transport systems, periplasmic component : braC |
| 163498557 | 0.00116 | COG3181 | K07795 | Mmwyl1_1799 | uncharacterized BCR : putative tricarboxylic transport membrane protein |
| 163497259 | 0.00107 | COG0747 | | CMM_2185 | ABC-type dipeptide/oligopeptide/nickel transport systems, periplasmic components |
| 163296806 | 0.00075 | COG0834 | K02030 | SAR11_1068 | ABC-type amino acid transport system, periplasmic component: pheC; cyclohexadienyl dehydratase; polar amino acid transport system substrate-binding protein |
| 163277703 | 0.00071 | COG0055 | K02112 | Acel_0653 | F0F1-type ATP synthase beta subunit |
| 163152101* <i>f</i> | 0.00066 | COG0459 | K04077 | SAR11_0162 | chaperonin GroEL (HSP60 family) |
| 163296936 <i>a</i> | 0.00055 | COG0050 | K02358 | SAR11_1130 | GTPases - translation elongation factors : tufB |
| 163117667 | 0.00043 | COG0174 | K01915 | SAR11_0747 | glutamine synthase : glnA |
| 163154554 <i>a</i> | 0.00042 | COG0050 | K02358 | Tfu_2648 | GTPases - translation elongation factors : tuf |
| 163450946 | 0.00036 | COG0683 | K01999 | SAR11_1346 | ABC-type branched-chain amino acid transport systems, periplasmic component : livJ |
| 163208382 | 0.00035 | COG0174 | K01915 | CMM_1636 | glutamine synthase : glnA |
| 163135037 | 0.00031 | COG2133 | K00540 | RspH17025_1771 | glucose/sorbose dehydrogenases |
| 163150509 | 0.00028 | COG0737 | | Haur_2906 | 5'-nucleotidase/2',3'-cyclic phosphodiesterase and related esterases |
| 163195194 | 0.00028 | COG0086 | K03046 | Lxx20630 | DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn) : rpoC |

Continued on next page

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--|---------|--------------|------------------|------------|--|
| 163169240* | 0.00027 | COG0330 | K04088 | SAR11_0008 | membrane protease subunits, stomatin/prohibitin homologs : hflK |
| 163135897 | 0.00014 | COG1185 | K00962 | SAR11_0392 | polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) : pnp; polynucleotide phosphorylase/polyadenylase |
| 11.5 m – KEGG and NR annotated proteins | | | | | |
| 163498919g | 0.04693 | BAF91544 | | | major capsid protein [uncultured Myoviridae] |
| 163303017h | 0.03153 | | GDI3673 | | hypothetical protein |
| 163496543i | 0.03095 | YP_001648158 | | | hypothetical protein [Ostreococcus virus OsV5] |
| 163312513 | 0.02334 | YP_002590925 | | | putative porin [Candidatus Pelagibacter sp. HTCC7211] |
| 163114028g | 0.02078 | YP_214669 | | | gp23 [Prochlorococcus phage P-SSM4] |
| 163447324i | 0.01947 | YP_001648266 | | | hypothetical protein OsV5_190f [Ostreococcus virus OsV5] |
| 163104039j | 0.01469 | YP_001498525 | AR158_C444L | | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 163299338 | 0.01058 | | Sputw3181_2479 | | phage major capsid protein, HK97 family |
| 163431599i | 0.01022 | YP_001648266 | OsV5_190f | | hypothetical protein [Ostreococcus virus OsV5] |
| 163486997k | 0.00971 | | SG1188 | | hypothetical protein |
| 163200650 | 0.00948 | | mlr8524 | | phage major capsid protein, GP36 |
| 163146271*l | 0.00930 | | BBta_5785 | | putative phage major head protein |
| 163466160 | 0.00921 | | BBta_5785 | | putative phage major head protein |
| 163111996h | 0.00849 | | GDI3673 | | hypothetical protein |
| 163277976 | 0.00767 | | Neut_1469 | | phage major capsid protein, HK97 family protein |
| 163276037 | 0.00762 | | mma_2202 | | hypothetical protein |
| 163114610i | 0.00717 | A7U6E7 | | | putative major capsid protein [Chrysochromulina ericina virus] |
| 163191828 | 0.00696 | ZP_03701413 | Flav3CDRAFT_1333 | | hypothetical protein [Flavobacteria bacterium MS024-3C] |
| 163121725* | 0.00678 | YP_001648124 | OsV5_047f | | hypothetical protein [Ostreococcus virus OsV5] |
| 163383538* | 0.00611 | | Neut_1469 | | phage major capsid protein, HK97 family protein |
| 163125121g | 0.00600 | YP_214669 | | | gp23 [Prochlorococcus phage P-SSM4] |
| 163161438 | 0.00539 | ABW90951 | | | gp23 major capsid protein [uncultured Myoviridae] |
| 163404994m | 0.00493 | | Haur_0657 | | hypothetical protein |
| 163115173*n | 0.00467 | YP_001648182 | OsV5_105r | | hypothetical protein [Ostreococcus virus OsV5] |
| 163253040 | 0.00435 | | Bpro_3745 | | hypothetical protein |
| 163519031 | 0.00433 | YP_002276820 | Gdia_2460 | | hypothetical protein [Gluconacetobacter diazotrophicus PA1 5] |
| 163228214o | 0.00413 | | Daci_1946 | | putative phage major head protein |
| 163291274h | 0.00407 | | GDI3673 | | hypothetical protein |
| 163206524*i | 0.00403 | YP_001648266 | OsV5_190f | | hypothetical protein [Ostreococcus virus OsV5] |

Continued on next page

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------------|---------|--------------|----|----------------|---|
| 163514201 | 0.00400 | | | MAB_1788 | bacteriophage protein |
| 163498539 | 0.00399 | | | SAR11_1290 | TRAP-type bacterial extracellular solute-binding protein |
| 163432666 | 0.00396 | | | Oter_3421 | hypothetical protein |
| 163480087 | 0.00392 | | | APEC01_525 | hypothetical protein |
| 163187860 | 0.00390 | | | GDI3673 | hypothetical protein |
| 163529078 | 0.00375 | | | HM1_2880 | phage major capsid protein, hk97 family |
| 163526011 <i>i</i> | 0.00371 | A7U6E9 | | | putative major capsid protein [Pyramimonas orientalis virus] |
| 163180584 | 0.00369 | BAE06835 | | | hypothetical major capsid protein [Heterosigma akashiwo virus 01] |
| 163459594 <i>i</i> | 0.00364 | BAE06835 | | | hypothetical major capsid protein [Heterosigma akashiwo virus 01] |
| 163503842 <i>h</i> | 0.00357 | | | GDI3673 | hypothetical protein |
| 163495193 | 0.00351 | | | MAB_1788 | bacteriophage protein |
| 163385358 <i>p</i> | 0.00324 | | | GDI3673 | hypothetical protein |
| 163489449 | 0.00313 | | | SG1188 | hypothetical protein |
| 163472957 | 0.00310 | | | Asuc_1240 | phage major capsid protein, HK97 family |
| 163131623 | 0.00298 | YP_195142 | | | major capsid protein gp23 [Synechococcus phage S-PM2] |
| 163118697 <i>i</i> | 0.00289 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 163420549 | 0.00284 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 163142179 | 0.00279 | ZP_03643684 | | BACCOPRO_02057 | hypothetical protein [Bacteroides coprophilus DSM 18228] |
| 163250350 <i>q</i> | 0.00277 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 163541257 | 0.00271 | | | Swit_4452 | hypothetical protein |
| 163478791 | 0.00261 | | | amb4267 | hypothetical protein |
| 163452556 | 0.00245 | | | APEC01_525 | hypothetical protein |
| 163507581 | 0.00237 | | | Cthe_2848 | phage major capsid protein, HK97 |
| 163409546* | 0.00234 | YP_001648301 | | OsV5_225r | hypothetical protein [Ostreococcus virus OsV5] |
| 163412911 | 0.00233 | ZP_03724502 | | ObacDRAFT_9001 | hypothetical protein [Opitutaceae bacterium TAV2] |
| 163544869 | 0.00230 | | | Smed_1334 | phage major capsid protein, HK97 family |
| 163323331 | 0.00222 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 163235228* | 0.00221 | | | BBta_5785 | putative phage major head protein |
| 163494515 <i>r</i> | 0.00217 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 163372194 | 0.00216 | ZP_01017474 | | | major capsid protein, HK97 family protein [Parvularcula bermudensis HTCC2503] |
| 163252031 <i>i</i> | 0.00210 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 163390078 <i>p</i> | 0.00210 | | | GDI3673 | hypothetical protein |
| 163290252 <i>m</i> | 0.00201 | | | Haur_0657 | hypothetical protein |

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|---------|--------------|--------|-------------|--|
| 163157042k | 0.00196 | | | SG1188 | hypothetical protein |
| 163199564i | 0.00192 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163490373h | 0.00191 | | | GDI3673 | hypothetical protein |
| 163445182 | 0.00189 | | | Acid_4111 | hypothetical protein |
| 163229276i | 0.00187 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163485571i | 0.00185 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 163499762 | 0.00183 | | | BBta_5785 | putative phage major head protein |
| 163227690i | 0.00181 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 163168692 | 0.00179 | | | M446_5960 | hypothetical protein |
| 163109620* | 0.00175 | YP_001648249 | | OsV5_172f | hypothetical protein [Ostreococcus virus OsV5] |
| 163105813 | 0.00173 | | | APEC01_4044 | hypothetical protein |
| 163141843*i | 0.00172 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163117897h | 0.00163 | | | GDI3673 | hypothetical protein |
| 163173092 | 0.00163 | | | Pmen_3970 | phage major capsid protein, HK97 family |
| 163491889 | 0.00161 | | | CKO_01864 | hypothetical protein |
| 163453714h | 0.00159 | | | GDI3673 | hypothetical protein |
| 163161098q | 0.00157 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 163177212 | 0.00156 | | | BBta_6597 | putative peptidase S14, ClpP |
| 163352152 | 0.00154 | | | BDI_2873 | putative outer membrane protein, probably involved in nutrient binding |
| 163372026i | 0.00151 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163287382q | 0.00151 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 163379834h | 0.00151 | | | GDI3673 | hypothetical protein |
| 163124973 | 0.00148 | | | BAV1464 | major capsid protein |
| 163354170 | 0.00145 | YP_001648190 | | OsV5_113r | hypothetical protein [Ostreococcus virus OsV5] |
| 163115568*s | 0.00143 | YP_001648153 | | OsV5_076f | hypothetical protein [Ostreococcus virus OsV5] |
| 163410122s | 0.00140 | YP_001648315 | | OsV5_239r | hypothetical protein [Ostreococcus virus OsV5] |
| 163411861r | 0.00138 | | | LGAS_1485 | predicted phage phi-C31 GP36 major capsid-like protein |
| 163220019 | 0.00132 | ZP_00743477 | | RBTH_08297 | hypothetical protein [Bacillus thuringiensis serovar israelensis ATCC 35646] |
| 163134239* | 0.00132 | YP_001648234 | | OsV5_157f | hypothetical protein [Ostreococcus virus OsV5] |
| 163174626* | 0.00130 | | K06904 | BL0376 | hypothetical protein with similarity to putative maturation protease of prophage CP-9 33CE |
| 163154474 | 0.00125 | | | gll0198 | similar to bacteriopsin |
| 163467688l | 0.00115 | | | Daci_1946 | putative phage major head protein |
| 163256412 | 0.00111 | ZP_00743477 | | RBTH_08297 | hypothetical protein [Bacillus thuringiensis serovar israelensis ATCC 35646] |

Continued on next page

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------------|---------|--------------|--------|-------------------|--|
| 163389410 | 0.00111 | | | Oant_1504 | peptidase U35 phage prohead HK97 |
| 163248889 <i>i</i> | 0.00110 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163415470 <i>i</i> | 0.00104 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 163191410* | 0.00100 | YP_001648184 | | OsV5_107r | hypothetical protein [Ostreococcus virus OsV5] |
| 163142589 | 0.00093 | | | Pmen_3970 | phage major capsid protein, HK97 family |
| 163327003 <i>t</i> | 0.00088 | | K06907 | Dde_1889 | hypothetical protein; K06907 |
| 163211634*i | 0.00086 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163393172 <i>i</i> | 0.00075 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163161074 <i>t</i> | 0.00073 | | K06907 | Dde_1889 | hypothetical protein; K06907 |
| 163141653*i | 0.00070 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 163110018 | 0.00068 | YP_001294637 | | ORF044 | hypothetical protein [Pseudomonas phage PA11] |
| 163249021 <i>s</i> | 0.00067 | YP_001648315 | | OsV5_239r | hypothetical protein [Ostreococcus virus OsV5] |
| 163445294*j | 0.00062 | YP_001498525 | | AR158_C444L | hypothetical protein [Paramecium bursaria Chlorella virus AR158] |
| 163507277 | 0.00058 | | | CHLREDRAFT_18622B | hypothetical protein |
| 163298764 <i>i</i> | 0.00058 | YP_001648158 | | OsV5_081f | hypothetical protein [Ostreococcus virus OsV5] |
| 163298596 <i>i</i> | 0.00055 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 163335150 | 0.00053 | | | BAV1464 | major capsid protein |
| 163327023*q | 0.00046 | YP_214367 | | | T4-like major capsid protein [Prochlorococcus phage P-SSM2] |
| 163287366 <i>t</i> | 0.00038 | | K06907 | Sfum_3815 | phage tail sheath protein; K06907 |
| 163475851 <i>o</i> | 0.00028 | | | ZMO0387 | major head protein |
| 163195236*s | 0.00027 | YP_001648315 | | OsV5_239r | hypothetical protein [Ostreococcus virus OsV5] |
| 163184761 | 0.00027 | YP_001648134 | | OsV5_057f | hypothetical protein [Ostreococcus virus OsV5] |
| 163109424*n | 0.00023 | YP_001648182 | | OsV5_105r | hypothetical protein [Ostreococcus virus OsV5] |
| 163368976 | 0.00020 | YP_001648211 | | OsV5_134r | hypothetical protein [Ostreococcus virus OsV5] |
| 163306940 | 0.00018 | YP_001648185 | | OsV5_108r | hypothetical protein [Ostreococcus virus OsV5] |
| 163162936* | 0.00017 | YP_001648263 | | OsV5_187r | hypothetical protein [Ostreococcus virus OsV5] |
| 163151745 <i>s</i> | 0.00017 | YP_001648315 | | OsV5_239r | hypothetical protein [Ostreococcus virus OsV5] |

11.5 m – Proteins with no annotation

| | |
|--------------------|---------|
| 163171140 | 0.03385 |
| 163345623 <i>u</i> | 0.03193 |
| 163279609 | 0.01466 |
| 163534693 | 0.00797 |
| 163109584 | 0.00788 |

Continued on next page

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------------|---------|-----------|----|-------|---------------------------|
| 163251027 | 0.00668 | | | | |
| 163250059 <i>u</i> | 0.00668 | | | | |
| 163386750 <i>u</i> | 0.00650 | | | | |
| 163254426 | 0.00638 | | | | |
| 163129699 | 0.00610 | | | | |
| 163113296* | 0.00528 | | | | |
| 163129983 | 0.00511 | | | | |
| 163395912 | 0.00471 | | | | |
| 163346783 | 0.00446 | | | | |
| 163246177 <i>u</i> | 0.00435 | | | | |
| 163303354 | 0.00395 | | | | |
| 163477113 | 0.00385 | | | | |
| 163419181 | 0.00349 | | | | |
| 163431790 <i>v</i> | 0.00331 | | | | |
| 163502200 | 0.00289 | | | | |
| 163456165 | 0.00285 | | | | |
| 163397872 | 0.00285 | | | | |
| 163490375 | 0.00278 | | | | |
| 163502202 | 0.00250 | | | | |
| 163453476* | 0.00247 | | | | |
| 163503840 | 0.00234 | | | | |
| 163187858 | 0.00225 | | | | |
| 163224309 <i>v</i> | 0.00211 | | | | |
| 163117895 | 0.00204 | | | | |
| 163285151 <i>w</i> | 0.00191 | | | | |
| 163511023 <i>u</i> | 0.00182 | | | | |
| 163156214 | 0.00160 | | | | |
| 163311655 | 0.00157 | | | | |
| 163286408 <i>w</i> | 0.00138 | | | | |
| 163254680 | 0.00138 | | | | |
| 163439545* | 0.00133 | | | | |
| 163123675 | 0.00131 | | | | |
| 163199154 <i>u</i> | 0.00126 | | | | |

Table B.2 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|----|-------|---------------------------|
| 163129981 | 0.00097 | | | | |
| 163110772* | 0.00096 | | | | |
| 163211312 | 0.00094 | | | | |
| 163207714 | 0.00092 | | | | |
| 163342613 | 0.00074 | | | | |
| 163217867* | 0.00071 | | | | |
| 163110016 | 0.00069 | | | | |
| 163117805 | 0.00063 | | | | |
| 163280533 | 0.00059 | | | | |
| 163168938 | 0.00059 | | | | |
| 163320197 | 0.00040 | | | | |

Table B.3: Proteins identified in the Ace Lake 12.7 m sample 0.1 μ m size-fraction metaproteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (a-z, aa-pp) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

| Gene ID | NSA | COG/NR ID | KO | Locus | 12.7 m – COG annotated proteins |
|-------------|---------|-----------|--------|-----------|---|
| | | | | | COG : KEGG/NR description |
| 165547755* | 0.01035 | COG0539 | K02945 | Cvib_1514 | rpsA; 30S ribosomal protein S1 |
| 165526280 | 0.00768 | COG0181 | K01749 | Cvib_1245 | porphobilinogen deaminase : hydroxymethylbilane synthase |
| 165511899 | 0.00756 | COG0516 | K00088 | Cvib_1056 | IMP dehydrogenase/GMP reductase |
| 165562959 | 0.00530 | COG1104 | K04487 | Cvib_0301 | cysteine sulfinate desulfurinase/cysteine desulfurase and related enzymes : aminotransferase, class V |
| 165514395* | 0.00452 | COG0674 | K00174 | Cvib_1597 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha sub-unit |
| 165502373 | 0.00380 | COG0129 | K01687 | Cvib_1169 | dihydroxy-acid dehydratase |
| 165514465* | 0.00340 | COG0054 | K00794 | Cvib_1632 | riboflavin synthase beta-chain |
| 165525758 | 0.00270 | COG1862 | K03210 | Cvib_0223 | preprotein translocase subunit YajC |
| 165514421* | 0.00243 | COG0250 | K02601 | Cvib_1610 | transcription antitermination protein NusG |
| 165526166 | 0.00209 | COG0413 | K00606 | Cvib_0725 | ketopantoate hydroxymethyltransferase : panB; 3-methyl-2-oxobutanoate hydroxymethyltransferase |
| 165562961 | 0.00203 | COG0031 | K01738 | Cvib_0300 | cysteine synthase |
| 165526282 | 0.00189 | COG1587 | K01719 | Cvib_1246 | uroporphyrinogen-III synthase |
| 165514409* | 0.00173 | COG0086 | K03046 | Cvib_1604 | DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn) |
| 165514577* | 0.00162 | COG1778 | K03270 | Cvib_1694 | uncharacterized proteins of HAD superfamily, CMP-Neu5Ac homologs : 3-deoxy-D-manno-octulosonate 8-phosphatase, YrbI family; (KDO 8-P phosphatase) |
| 165514581*a | 0.00158 | COG0542 | | Cvib_1696 | ATPases with chaperone activity, ATP-binding subunit : AAA-2 domain protein |
| 165536856 | 0.00155 | COG0157 | K00767 | Cvib_0335 | nicotinate-nucleotide pyrophosphorylase [carboxylating] |
| 165547841 | 0.00146 | COG0493 | K00266 | Cvib_1478 | NADPH-dependent glutamate synthase beta chain and related oxidoreductases |
| 165514651* | 0.00139 | COG0797 | K03642 | Cvib_1727 | lipoproteins : rare lipoprotein A |
| 165525906 | 0.00137 | COG0750 | K01417 | Cvib_0137 | predicted membrane-associated Zn-dependent proteases 1 |
| 165501975* | 0.00135 | COG0740 | K01358 | Cvib_0441 | protease subunit of ATP-dependent Clp proteases |
| 165505943 | 0.00135 | COG0543 | | Cvib_0839 | 2-polypropenylphenol hydroxylase and related flavodoxin oxidoreductases : oxidoreductase FAD/NAD(P)-binding domain protein |
| 165526296 | 0.00133 | COG0082 | K01736 | Cvib_1253 | chorismate synthase |
| 165547993 | 0.00132 | COG0497 | K03631 | Cvib_1402 | ATPases involved in DNA repair : DNA repair protein RecN |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|------------|--|
| 165547777 | 0.00126 | COG0008 | K01885 | Cvib_1503 | glutamyl-tRNA synthetase |
| 165553075 | 0.00125 | COG1158 | K03628 | Cvib_1537 | transcription termination factor : Rho |
| 165511737 | 0.00121 | COG1726 | K03615 | Cvib_0797 | Na ⁺ -transporting NADH:ubiquinone oxidoreductase alpha subunit : electron transport complex, RnfABCDGE type, C subunit |
| 165526284 | 0.00118 | COG0483 | K01092 | Cvib_1247 | archaeal fructose-1,6-bisphosphatase and related enzymes of inositol monophosphatase family |
| 165525808 | 0.00116 | COG0331 | K00645 | Cvib_0199 | (acyl-carrier-protein) S-malonyltransferase |
| 165550953 | 0.00116 | COG1022 | K01897 | Cvib_0930 | long-chain acyl-CoA synthetases (AMP-forming) : AMP-dependent synthetase and ligase |
| 165502369 | 0.00105 | COG0440 | K01653 | Cvib_1171 | acetolactate synthase, small subunit |
| 165526250 | 0.00104 | COG1522 | | Cvib_1231 | transcriptional regulators : AsnC family |
| 165525708 | 0.00102 | COG0089 | K02892 | Cvib_0248 | rplW; 50S ribosomal protein L23 |
| 165502825* | 0.00102 | COG0341 | K03074 | Cvib_0011 | secF; preprotein translocase subunit SecF |
| 165502225 | 0.00101 | COG0557 | K01147 | Cvib_0574 | Exoribonucleases : RNAse R; exoribonuclease II |
| 165514389* | 0.00101 | COG0446 | | Cvib_1594 | uncharacterized NAD(FAD)-dependent dehydrogenases : FAD-dependent pyridine nucleotide-disulphide oxidoreductase |
| 165525802 | 0.00100 | COG0333 | K02911 | Cvib_0202 | rpmF; 50S ribosomal protein L32 |
| 165525664 | 0.00098 | COG0100 | K02948 | Cvib_0271 | 30S ribosomal protein S11 |
| 165511903 | 0.00095 | COG1240 | K03404 | Cvib_1058 | Mg-chelatase subunit ChII : protoporphyrin IX magnesium-chelatase |
| 165514579* | 0.00094 | COG2877 | K01627 | Cvib_1695 | 3-Deoxy-D-manno-octulose-2-phosphate synthase |
| 165525806 | 0.00094 | COG0332 | K00648 | Cvib_0200 | 3-oxoacyl-(acyl carrier protein) synthase III |
| 165519368 | 0.00094 | COG1192 | K03496 | Cvib_0388 | ATPases involved in chromosome partitioning |
| 165547931 | 0.00093 | COG0217 | | Cvib_1432 | uncharacterized ACR : hypothetical protein |
| 165536866 | 0.00091 | COG0468 | K03553 | Cvib_0340 | RecA/RadA recombinase |
| 165514413* | 0.00086 | COG0222 | K02935 | Cvib_1606 | rplL; 50S ribosomal protein L7/L12 |
| 165547847 | 0.00086 | COG0106 | K01814 | Cvib_1475 | phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase |
| 165525912 | 0.00085 | COG0778 | | Cvib_0134 | nitroreductase |
| 165547827 | 0.00082 | COG1136 | K02003 | Cvib_1485 | ABC-type transport systems, involved in lipoprotein release, ATPase components |
| 165547771 | 0.00080 | COG0776 | K05788 | Cvib_1506 | bacterial nucleoid DNA-binding protein : histone family protein DNA-binding protein; integration host factor subunit beta |
| 165562901 | 0.00080 | COG0003 | K01551 | Cvib_0328 | predicted ATPase involved in chromosome partitioning : arsenite-activated ATPase ArsA |
| 165530868 | 0.00079 | COG2089 | K01654 | fnu:FN1684 | sialic acid synthase : N-acetylneuraminate synthase |
| 165562905 | 0.00078 | COG0629 | K03111 | Cvib_0326 | single-strand DNA-binding protein |
| 165502829* | 0.00077 | COG0446 | K00540 | Cvib_0009 | uncharacterized NAD(FAD)-dependent dehydrogenases : sulfide dehydrogenase (flavocytochrome), flavoprotein subunit |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|---|
| 165511991 | 0.00076 | COG1418 | K06950 | Cvib_1112 | predicted HD superfamily hydrolase : hypothetical protein |
| 165547765 | 0.00075 | COG0503 | K00759 | Cvib_1509 | adenine/guanine phosphoribosyltransferases and related PRPP-binding proteins |
| 165515181 | 0.00075 | COG0711 | K02109 | Cvib_1741 | F0F1-type ATP synthase b subunit |
| 165502313 | 0.00075 | COG0209 | K00525 | Cvib_1199 | ribonucleotide-diphosphate reductase subunit alpha |
| 165525750a | 0.00074 | COG0542 | | Cvib_0227 | ATPases with chaperone activity, ATP-binding subunit : AAA ATPase, central domain protein |
| 165525840 | 0.00074 | COG0360 | K02990 | Cvib_0181 | rpsF; 30S ribosomal protein S6 |
| 165547781 | 0.00073 | COG0723 | K02636 | Cvib_1501 | Rieske Fe-S protein : plastoquinol–plastocyanin reductase; cytochrome b6-f complex iron-sulfur subunit |
| 165548025 | 0.00073 | COG0058 | K00688 | Cvib_1386 | alpha-glucan phosphorylase |
| 165511773 | 0.00072 | COG0524 | | Cvib_0779 | sugar kinases, ribokinase family : PfkB domain protein |
| 165514433* | 0.00071 | COG0480 | K02355 | Cvib_1616 | translation elongation and release factors (GTPases) : fusA; elongation factor G |
| 165526086 | 0.00071 | COG2089 | K01654 | Cvib_1025 | sialic acid synthase : N-acetylneuraminate synthase |
| 165502143 | 0.00070 | COG3040 | K03098 | Cvib_0516 | bacterial lipocalin |
| 165547991 | 0.00070 | COG0329 | K01714 | Cvib_1403 | dihydrodipicolinate synthase/N-acetylneuraminate lyase |
| 165536870 | 0.00069 | COG0136 | K00133 | Cvib_0342 | aspartate semialdehyde dehydrogenase |
| 165502129 | 0.00067 | COG0158 | K03841 | Cvib_0509 | fructose-1,6-bisphosphatase |
| 165525834 | 0.00067 | COG0359 | K02939 | Cvib_0184 | rplI; 50S ribosomal protein L9 |
| 165547783 | 0.00065 | COG1290 | K00412 | Cvib_1500 | cytochrome b subunit of the bc complex : cytochrome b/b6, N-terminal domain protein; ubiquinol-cytochrome c reductase |
| 165553031 | 0.00065 | COG0184 | K02956 | Cvib_1557 | rpsO; 30S ribosomal protein S15 |
| 165502097 | 0.00065 | COG0127 | K01516 | Cvib_0493 | non-canonical purine NTP pyrophosphatase, RdgB/HAM1 family |
| 165514535* | 0.00064 | COG0366 | | Cvib_1672 | glycosidases : trehalose synthase |
| 165536860 | 0.00064 | COG0544 | K03545 | Cvib_0337 | FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor) |
| 165519356 | 0.00060 | COG1410 | K00548 | Cvib_0382 | methionine synthase I, cobalamin-binding domain : metH; 5-methyltetrahydrofolate–homocysteine methyltransferase |
| 165525900 | 0.00058 | COG0284 | K01591 | Cvib_0140 | orotidine 5'-phosphate decarboxylase |
| 165547983 | 0.00057 | COG0674 | K03737 | Cvib_1407 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha sub-unit |
| 165547901 | 0.00057 | COG3155 | | Cvib_1447 | uncharacterized sigma cross-reacting protein 27A (ES1 or KNP-I alpha protein) : isoprenoid biosynthesis protein with amidotransferase-like domain |
| 165562941 | 0.00057 | COG2319 | | Cvib_0309 | WD-40 repeat protein |
| 165526078 | 0.00056 | COG0589 | | Cvib_1021 | universal stress protein UspA and related nucleotide-binding proteins |
| 165516491 | 0.00056 | COG0760 | | Cvib_1572 | parvulin-like peptidyl-prolyl isomerase : PpiC-type peptidyl-prolyl cis-trans isomerase |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|---|
| 165562981 | 0.00053 | COG1043 | K00677 | Cvib_0290 | acyl-[acyl carrier protein]-UDP-N-acetylglucosamine O-acyltransferase |
| 165502845* | 0.00052 | COG0706 | K03217 | Cvib_1771 | preprotein translocase subunit YidC |
| 165525718 | 0.00051 | COG0480 | K02355 | Cvib_0243 | translation elongation and release factors (GTPases) : fusA; elongation factor G |
| 165526290 | 0.00051 | COG0511 | | Cvib_1250 | biotin carboxyl carrier protein : biotin/lipoyl attachment domain-containing protein |
| 165557791 | 0.00051 | COG0074 | | Cvib_0866 | succinyl-CoA synthetase alpha subunit : ATP citrate lyase subunit 2 |
| 165547973 | 0.00050 | COG1493 | K06023 | Cvib_1412 | serine kinase of the HPr protein, regulates carbohydrate metabolism |
| 165502247 | 0.00050 | COG0234 | K04078 | Plut_0541 | groES; co-chaperonin GroES |
| 165514671* | 0.00050 | COG0636 | K02110 | Plut_2097 | FOF1-type ATP synthase c subunit/Archaeal/vacuolar-type H+-ATPase subunit K : ATP synthase F0, C subunit |
| 165502267 | 0.00050 | COG1077 | K03569 | Cvib_0595 | HSP70 class molecular chaperones involved in cell morphogenesis : cell shape determining protein, MreB/Mrl family |
| 165525868 | 0.00049 | COG0188 | K02469 | Cvib_0172 | DNA gyrase subunit A |
| 165514647* | 0.00048 | COG0167 | K00226 | Cvib_1724 | dihydroorotate dehydrogenase 2 |
| 165509259* | 0.00046 | COG0793 | K03797 | Cvib_0018 | periplasmic protease : carboxyl-terminal protease |
| 165548143 | 0.00046 | COG0261 | K02888 | Cvib_1329 | rplU; 50S ribosomal protein L21 |
| 165514553* | 0.00045 | COG0001 | K01845 | Cvib_1681 | glutamate-1-semialdehyde 2,1-aminomutase |
| 165525814 | 0.00044 | COG0304 | K09458 | Cvib_0196 | 3-oxoacyl-[acyl-carrier-protein] synthase II |
| 165547833 | 0.00043 | COG2226 | K03183 | Cvib_1482 | methylase involved in ubiquinone/menaquinone biosynthesis : demethylmenaquinone methyl-transferase |
| 165547919 | 0.00042 | COG0036 | K01783 | Cvib_1438 | ribulose-5-phosphate 3-epimerase |
| 165525676 | 0.00042 | COG1841 | K02907 | Cvib_0264 | rpmD; 50S ribosomal protein L30 |
| 165502421 | 0.00042 | COG0443 | K04043 | Cvib_1158 | chaperone protein DnaK |
| 165525692 | 0.00042 | COG0093 | K02874 | Cvib_0256 | rplN; 50S ribosomal protein L14 |
| 165525970 | 0.00042 | COG1220 | K03667 | Cvib_0959 | ATP-dependent protease, ATPase subunit : hslU |
| 165525828 | 0.00041 | COG0292 | K02887 | Cvib_0187 | rplT; 50S ribosomal protein L20 |
| 165502239 | 0.00040 | COG0596 | | Cvib_0581 | predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily) |
| 165525710 | 0.00039 | COG0088 | K02926 | Cvib_0247 | rplD; 50S ribosomal protein L4 |
| 165502125 | 0.00039 | COG2838 | K00031 | Cvib_0507 | monomeric isocitrate dehydrogenase |
| 165526072 | 0.00039 | COG1038 | K01571 | Cvib_1018 | pyruvate carboxylase, C-terminal domain/subunit : biotin/lipoyl attachment domain-containing protein; oxaloacetate decarboxylase, alpha subunit |
| 165562947 | 0.00038 | COG0040 | K00765 | Cvib_0307 | hisG; ATP phosphoribosyltransferase |
| 165519380 | 0.00038 | COG0447 | K01661 | Cvib_0394 | dihydroxynaphthoic acid synthase |
| 165514407* | 0.00038 | COG0439 | K01961 | Cvib_1603 | acetyl-CoA carboxylase, biotin carboxylase |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|---|
| 165547843 | 0.00038 | COG0543 | K00528 | Cvib_1477 | 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases : ferredoxin–NADP(+) reductase subunit alpha |
| 165502249 | 0.00038 | COG0459 | K04077 | Cvib_0586 | chaperonin GroEL (HSP60 family) |
| 165525978 | 0.00038 | COG0190 | K00288 | Cvib_0963 | methenyltetrahydrofolate cyclohydrolase (NADP+) |
| 165548015 | 0.00037 | COG0115 | K00826 | Cvib_1391 | 4-amino-4-deoxychorismate lyase : branched-chain amino acid aminotransferase |
| 165526070 | 0.00037 | COG1883 | K01572 | Cvib_1017 | Na ⁺ -transporting methylmalonyl-CoA/oxaloacetate decarboxylase, beta subunit |
| 165502323 | 0.00037 | COG2406 | K03594 | Cvib_1194 | uncharacterized ACR : ferritin, Dps family protein |
| 165502157 | 0.00036 | COG0182 | K08963 | Cvib_0532 | translation initiation factor 2B subunit I family (IF-2BI); methylthioribose-1-phosphate isomerase |
| 165502159 | 0.00036 | COG0005 | K03783 | Cvib_0533 | purine nucleoside phosphorylase |
| 165526084 | 0.00035 | COG0326 | K04079 | Cvib_1024 | heat shock protein 90; molecular chaperone HtpG |
| 165547741* | 0.00034 | COG1109 | K01840 | Cvib_1521 | phosphoglucomutase |
| 165547943 | 0.00033 | COG0365 | K01895 | Cvib_1426 | acyl-coenzyme A synthetases/AMP-(fatty) acid ligases : acetyl-coenzyme A synthetase |
| 165547811 | 0.00033 | COG1118 | K02017 | Cvib_1494 | ABC-type sulfate/molybdate transport systems, ATPase component |
| 165501997* | 0.00032 | COG0404 | K00605 | Cvib_0451 | glycine cleavage system T protein (aminomethyltransferase) |
| 165548139 | 0.00032 | COG0039 | K00026 | Cvib_1331 | malate dehydrogenase |
| 165547815 | 0.00031 | COG0725 | K02020 | Cvib_1492 | ABC-type molybdate transport system, periplasmic component |
| 165526254 | 0.00031 | COG3349 | K00514 | Cvib_1233 | uncharacterized ACR : zeta-carotene desaturase |
| 165519338* | 0.00031 | COG1023 | K00033 | Cvib_0374 | 6-phosphogluconate dehydrogenase, family 2 |
| 165525774 | 0.00031 | COG0315 | K03637 | Cvib_0215 | moaC; bifunctional molybdenum cofactor biosynthesis protein C/molybdopterin-binding protein |
| 165514597* | 0.00031 | COG0248 | | Cvib_1704 | Ppx/GppA phosphatase |
| 165505937 | 0.00030 | COG1908 | | Cvib_0842 | coenzyme F420-reducing hydrogenase, delta subunit : methyl-viologen-reducing hydrogenase, delta subunit |
| 165505095 | 0.00030 | COG1192 | | Cvib_0908 | ATPases involved in chromosome partitioning : cobyric acid a,c-diamide synthase |
| 165547897 | 0.00030 | COG0178 | K03701 | Cvib_1449 | excinuclease ABC subunit A |
| 165553023 | 0.00030 | COG0532 | K02519 | Cvib_1561 | translation initiation factor 2 (GTPase) : infB |
| 165553083 | 0.00029 | COG3637 | | Cvib_1533 | opacity protein and related surface antigens : porin |
| 165562903 | 0.00029 | COG0543 | K02823 | Cvib_0327 | 2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases : dihydroorotate oxidase B, electron transfer subunit |
| 165547817 | 0.00029 | COG0157 | K03813 | Cvib_1491 | nicotinate-nucleotide pyrophosphorylase : ModD protein; molybdenum transport protein |
| 165511803 | 0.00028 | COG0152 | K01923 | Cvib_0763 | phosphoribosylaminoimidazole-succinocarboxamide synthase |
| 165502273 | 0.00028 | COG1049 | K01682 | Cvib_0598 | aconitase B: bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|--|
| 165525716 | 0.00028 | COG0050 | K02358 | Cvib_0244 | GTPases - translation elongation factors : tuf |
| 165505137* | 0.00027 | COG0724 | | Cvib_0890 | RNA-binding proteins (RRM domain) |
| 165511799 | 0.00027 | COG0021 | K00615 | Cvib_0765 | transketolase subunit A |
| 165502001* | 0.00027 | COG0694 | K07400 | Cvib_0453 | thioredoxin-like proteins and domains : nitrogen-fixing NifU domain protein |
| 165525700 | 0.00027 | COG0092 | K02982 | Cvib_0252 | rpsC; 30S ribosomal protein S3 |
| 165548043 | 0.00027 | COG0848 | | Cvib_1377 | biopolymer transport protein ExbD/TolR |
| 165502837* | 0.00027 | COG0426 | | Cvib_0005 | uncharacterized flavoproteins : beta-lactamase domain protein |
| 165502823* | 0.00027 | COG0760 | K03771 | Cvib_0012 | parvulin-like peptidyl-prolyl isomerase : PpiC-type peptidyl-prolyl cis-trans isomerase |
| 165525682 | 0.00026 | COG0097 | K02933 | Cvib_0261 | rplF; 50S ribosomal protein L6 |
| 165509275* | 0.00026 | COG0055 | K02112 | Cvib_0025 | F0F1 ATP synthase subunit beta |
| 165514365a | 0.00026 | COG0542 | K03696 | Cvib_1580 | ATPases with chaperone activity, ATP-binding subunit : AAA-2 domain protein; ATP-dependent Clp protease |
| 165525766 | 0.00026 | COG3245 | | Cvib_0219 | cytochrome c5 |
| 165514371* | 0.00026 | COG0451 | | Cvib_1582 | nucleoside-diphosphate-sugar epimerases : NAD-dependent epimerase/dehydratase |
| 165525810 | 0.00026 | COG1028 | K00059 | Cvib_0198 | dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) : 3-oxoacyl-[acyl-carrier-protein] reductase |
| 165502383 | 0.00025 | COG0823 | K03641 | Cvib_1164 | periplasmic component of the Tol biopolymer transport system : WD40 domain protein beta propeller; TolB protein |
| 165557763 | 0.00025 | COG0137 | K01940 | Cvib_0882 | argininosuccinate synthase |
| 165502177 | 0.00024 | COG0591 | | Cvib_0542 | Na+/proline, Na+/panthothenate symporters and related permeases |
| 165525720 | 0.00024 | COG0049 | K02992 | Cvib_0242 | 30S ribosomal protein S7 |
| 165553085 | 0.00024 | COG0729 | K07277 | Cvib_1532 | predicted outer membrane protein : surface antigen (D15) |
| 165525770 | 0.00024 | COG0345 | K00286 | Cvib_0217 | pyrroline-5-carboxylate reductase |
| 165514453* | 0.00023 | COG0527 | K00928 | Cvib_1626 | aspartokinases |
| 165509301* | 0.00023 | COG2920 | K00396 | Cvib_0038 | sulfite reductase, gamma subunit : DsrC family protein |
| 165514489* | 0.00023 | COG1538 | | Cvib_1644 | outer membrane protein : outer membrane efflux protein |
| 165562971 | 0.00022 | COG0105 | K00940 | Cvib_0295 | nucleoside diphosphate kinase |
| 165562957 | 0.00022 | COG0822 | K04488 | Cvib_0302 | NifU homologs involved in Fe-S cluster formation |
| 165501969* | 0.00022 | COG1752 | | Cvib_0421 | predicted esterase of the alpha-beta hydrolase superfamily : surface antigen (D15) |
| 165526170 | 0.00022 | COG1832 | K06929 | Cvib_0723 | predicted CoA-binding protein |
| 165514455* | 0.00021 | COG0224 | K02115 | Cvib_1627 | ATP synthase F1, gamma subunit |
| 165562919 | 0.00021 | COG1629 | K02014 | Plut_0256 | outer membrane receptor proteins, mostly Fe transport : ferric siderophore receptor, putative, TonB receptor family |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|--|
| 16550965 | 0.00020 | COG0541 | K03106 | Cvib_0936 | signal recognition particle GTPase : subunit FFH/SRP54 (SRP54) |
| 165502821* | 0.00020 | COG0187 | K02470 | Cvib_0013 | DNA gyrase subunit B |
| 165509279* | 0.00020 | COG1274 | K01596 | Cvib_0027 | phosphoenolpyruvate carboxykinase (GTP) |
| 165514471* | 0.00020 | COG0407 | K01599 | Cvib_1635 | uroporphyrinogen-III decarboxylase |
| 165505923 | 0.00020 | COG0653 | K03070 | Cvib_0853 | preprotein translocase subunit SecA (ATPase, RNA helicase) |
| 165526076 | 0.00020 | COG1951 | K01676 | Cvib_1020 | tartrate dehydratase alpha subunit/Fumarate hydratase class I, N-terminal domain |
| 165547989 | 0.00020 | COG0403 | K00282 | Cvib_1404 | glycine cleavage system protein P (pyridoxal-binding), N-terminal domain : glycine dehydrogenase subunit 1 |
| 165550943 | 0.00020 | COG0399 | | Cvib_0925 | predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis : DegT/DnrJ/EryC1/StrS aminotransferase |
| 165519308* | 0.00019 | COG1233 | | Cvib_0356 | phytoene dehydrogenase and related proteins : FAD dependent oxidoreductase |
| 165525704 | 0.00019 | COG0185 | K02965 | Cvib_0250 | rpsS; 30S ribosomal protein S19 |
| 165502209 | 0.00019 | COG0493 | K00266 | Cvib_0559 | NADPH-dependent glutamate synthase beta chain and related oxidoreductases : gltD |
| 165548183 | 0.00019 | COG0057 | K00134 | Cvib_1310 | glyceraldehyde-3-phosphate dehydrogenase |
| 165509303* | 0.00018 | COG2221 | K00396 | Cvib_0039 | oxidoreductase related to nitrite reductase : sulfite reductase, dissimilatory-type alpha subunit |
| 165502855* | 0.00018 | COG0126 | K00927 | Cvib_1766 | pgk; phosphoglycerate kinase |
| 165548189 | 0.00018 | COG0399 | | Cvib_1298 | predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis : DegT/DnrJ/EryC1/StrS aminotransferase |
| 165502083 | 0.00018 | COG0723 | K09879 | Cvib_0486 | Rieske Fe-S protein : isorenieratene synthase |
| 165502241 | 0.00018 | COG1959 | | Cvib_0582 | predicted transcriptional regulator : transcriptional regulator, BadM/Rrf2 family |
| 165505101 | 0.00018 | COG1744 | K07335 | Cvib_0906 | surface lipoprotein |
| 165562997 | 0.00018 | COG1348 | K04037 | Cvib_0283 | nitrogenase subunit NifH (ATPase) : chlL, bchL; protochlorophyllide reductase iron-sulfur ATP-binding protein |
| 165514479* | 0.00017 | COG0517 | | Plut_1996 | CBS domains |
| 165509257* | 0.00017 | COG2171 | K00674 | Cvib_0017 | tetrahydrodipicolinate N-succinyltransferase : dapD; 2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase |
| 165525898 | 0.00017 | COG0449 | K00820 | Cvib_0141 | glucosamine 6-phosphate synthetase, contains amidotransferase and phosphosugar isomerase domains : glutamine-fructose-6-phosphate transaminase |
| 165526040 | 0.00017 | COG0226 | K02040 | Cvib_0998 | phosphate binding protein; phosphate transport system substrate-binding protein |
| 165526106 | 0.00017 | COG0718 | K09747 | Cvib_1034 | uncharacterized BCR : conserved hypothetical protein 103 |
| 165525660 | 0.00017 | COG0202 | K03040 | Cvib_0273 | DNA-directed RNA polymerase subunit alpha |
| 165563007 | 0.00017 | COG0218 | K03978 | Cvib_0278 | yihA, ysxC, engB; GTPase EngB |
| 165526124 | 0.00016 | COG0589 | | Cvib_1041 | universal stress protein UspA and related nucleotide-binding proteins |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|--|
| 165526138 | 0.00016 | COG1554 | | Cvib_1047 | trehalose and maltose hydrolases (possible phosphorylases) : beta-phosphoglucomutase family hydrolase |
| 165525892 | 0.00016 | COG0724 | | Cvib_0162 | RNA-binding proteins (RRM domain) |
| 165514417* | 0.00016 | COG0081 | K02863 | Cvib_1608 | rplA; 50S ribosomal protein L1 |
| 165550985 | 0.00016 | COG0260 | K01255 | Cvib_0947 | leucyl aminopeptidase |
| 165514655* | 0.00016 | COG1899 | K00809 | Cvib_1729 | deoxyhypusine synthase |
| 165525670 | 0.00016 | COG0024 | K01265 | Cvib_0267 | methionine aminopeptidase : type I |
| 165526140 | 0.00015 | COG0809 | K07568 | Cvib_1048 | S-adenosylmethionine:tRNA-ribosyltransferase-isomerase (queuine synthetase) : queuosine biosynthesis protein |
| 165502015 | 0.00015 | COG0264 | K02357 | Cvib_0459 | elongation factor Ts : tsf |
| 165514419* | 0.00015 | COG0080 | K02867 | Plut_1966 | rplK; 50S ribosomal protein L11 |
| 165553033 | 0.00015 | COG1561 | | Cvib_1556 | uncharacterized stress-induced protein : hypothetical protein |
| 165519394 | 0.00015 | COG0372 | K01647 | Cvib_0401 | citrate synthase |
| 165548035 | 0.00015 | COG0112 | K00600 | Cvib_1381 | glycine hydroxymethyltransferase |
| 165525666 | 0.00014 | COG0099 | K02952 | Cvib_0270 | rpsM; 30S ribosomal protein S13 |
| 165502377 | 0.00014 | COG0811 | K03562 | Cvib_1167 | biopolymer transport proteins : MotA/TolQ/ExbB proton channel |
| 165519372 | 0.00014 | COG0289 | K00215 | Cvib_0390 | dihydridipicolinate reductase |
| 165526294 | 0.00014 | COG0113 | K01698 | Cvib_1252 | delta-aminolevulinic acid dehydratase |
| 165553057 | 0.00014 | COG0142 | K00795 | Cvib_1546 | geranyltransterase |
| 165511739 | 0.00014 | COG1805 | K03614 | Cvib_0796 | Na ⁺ -transporting NADH:ubiquinone oxidoreductase subunit 2 : electron transport complex, RnfABCDGE type, D subunit |
| 165553087 | 0.00014 | COG0020 | K00806 | Cvib_1531 | undecaprenyl pyrophosphate synthetase |
| 165502433 | 0.00013 | COG0633 | K08953 | Cvib_1151 | ferredoxin : chlorosome envelope protein J |
| 165526246 | 0.00013 | COG0174 | K01915 | Cvib_1230 | glutamine synthetase, catalytic region |
| 165502389 | 0.00013 | COG1729 | | Cvib_1161 | uncharacterized BCR : tetratricopeptide domain protein |
| 165562927 | 0.00012 | COG2265 | K03428 | Cvib_0317 | SAM-dependent methyltransferases related to tRNA (uracil-5)-methyltransferase : Mg-protoporphyrin IX methyl transferase |
| 165511967 | 0.00012 | COG1538 | | Cvib_1100 | outer membrane protein : outer membrane efflux protein |
| 165514411* | 0.00012 | COG0085 | K03043 | Cvib_1605 | rpoB; DNA-directed RNA polymerase subunit beta |
| 165502197 | 0.00012 | COG0045 | K01903 | Cvib_0553 | succinyl-CoA synthetase (ADP-forming) beta subunit |
| 165525686 | 0.00012 | COG0199 | K02954 | Plut_0194 | rpsN; 30S ribosomal protein S14 |
| 165548039 | 0.00012 | COG0811 | K03561 | Cvib_1379 | biopolymer transport proteins : MotA/TolQ/ExbB proton channel |
| 165509331* | 0.00012 | COG0007 | K02302 | Cvib_0053 | uroporphyrinogen-III C-methyltransferase |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|---|
| 165518169* | 0.00012 | COG2221 | K00396 | Cvib_0040 | oxidoreductase related to nitrite reductase : sulfite reductase, dissimilatory-type beta subunit |
| 165547929 | 0.00012 | COG0077 | K04518 | Cvib_1433 | prephenate dehydratase |
| 165557761 | 0.00012 | COG0165 | K01755 | Cvib_0883 | argininosuccinate lyase |
| 165511743 | 0.00012 | COG1347 | K03613 | Cvib_0794 | Na ⁺ -transporting NADH:ubiquinone oxidoreductase subunit 4 : SoxR-reducing system protein RsxE; electron transport complex protein RnfE |
| 165511765 | 0.00012 | COG0003 | K01551 | Cvib_0783 | predicted ATPase involved in chromosome partitioning : arsenite-activated ATPase ArsA |
| 165502673* | 0.00011 | COG1396 | | Plut_1890 | predicted transcriptional regulators: XRE family |
| 165505117 | 0.00011 | COG0698 | K01808 | Cvib_0896 | ribose 5-phosphate isomerase RpiB |
| 165526204 | 0.00011 | COG0019 | K01586 | Cvib_0705 | diaminopimelate decarboxylase |
| 165526060 | 0.00011 | COG0526 | K03671 | Cvib_1012 | thiol-disulfide isomerase and thioredoxins |
| 165501979* | 0.00010 | COG0568 | K03086 | Cvib_0443 | DNA-directed RNA polymerase sigma subunits (sigma70/sigma32) : RpoH |
| 165525702 | 0.00010 | COG0091 | K02890 | Cvib_0251 | rplV; 50S ribosomal protein L22 |
| 165562969 | 0.00010 | COG1225 | | Cvib_0296 | peroxiredoxin |
| 165548179 | 0.00010 | COG0484 | K03686 | Cvib_1312 | molecular chaperones (contain C-terminal Zn finger domain) : chaperone protein DnaJ |
| 165526058 | 0.00010 | COG0492 | K00384 | Cvib_1011 | thioredoxin reductase |
| 165514477* | 0.00010 | COG1053 | K00239 | Cvib_1638 | succinate dehydrogenase/fumarate reductase, flavoprotein subunits |
| 165514475* | 0.00010 | COG0479 | K00240 | Cvib_1637 | succinate dehydrogenase/fumarate reductase Fe-S protein : succinate dehydrogenase subunit B |
| 165514463* | 0.00010 | COG0204 | K00655 | Cvib_1631 | 1-acyl-sn-glycerol-3-phosphate acyltransferase |
| 165512005 | 0.00010 | COG0499 | K01251 | Cvib_1122 | S-adenosyl-L-homocysteine hydrolase; adenosylhomocysteinase |
| 165511973 | 0.00010 | COG2077 | K00435 | Cvib_1103 | peroxiredoxin : thiol peroxidase (atypical 2-Cys peroxiredoxin) |
| 165519288* | 0.00009 | COG0031 | K01697 | Cvib_0346 | cysteine synthase |
| 165502065 | 0.00009 | COG2177 | K09811 | Cvib_0478 | cell division protein FtsX |
| 165502349 | 0.00009 | COG1610 | K09117 | Cvib_1181 | uncharacterized ACR : GatB/YqeY domain protein |
| 165502361 | 0.00009 | COG0065 | K01703 | Cvib_1175 | homoaconitate hydratase family protein; K01703 3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit |
| 165526160 | 0.00009 | COG0047 | K01952 | Cvib_0728 | phosphoribosylformylglycinamide synthase I |
| 165526016 | 0.00009 | COG3118 | K05838 | Cvib_0982 | thioredoxin domain-containing protein |
| 165525690 | 0.00009 | COG0198 | K02895 | Plut_0192 | rplX; 50S ribosomal protein L24 |
| 165519322* | 0.00009 | COG2606 | | Cvib_0364 | uncharacterized ACR : YbaK/prolyl-tRNA synthetase associated region |
| 165502387 | 0.00009 | COG2885 | | Cvib_1162 | outer membrane protein and related peptidoglycan-associated (lipo)proteins : OmpA/MotB domain protein |
| 165525762 | 0.00009 | COG0075 | K00839 | Cvib_0221 | serine-pyruvate aminotransferase/archaeal aspartate aminotransferase : aminotransferase, class V |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|--|
| 165511763 | 0.00009 | COG0722 | K01626 | Cvib_0784 | 3-Deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase : phospho-2-dehydro-3-heoxyheptonate aldolase; 3-deoxy-7-phosphoheptulonate synthase |
| 165525764 | 0.00009 | COG3245 | | Cvib_0220 | cytochrome c5 |
| 165514427* | 0.00009 | COG0821 | K03526 | Cvib_1613 | essential bacterial protein, involved in density-dependent regulation of peptidoglycan biosynthesis : 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase |
| 165525908 | 0.00009 | COG0216 | K02835 | Cvib_0136 | prfA; peptide chain release factor 1 |
| 165526270 | 0.00009 | COG1217 | K06207 | Cvib_1242 | predicted membrane GTPase involved in stress response : GTP-binding protein TypA |
| 165526178 | 0.00008 | COG0845 | K02005 | Cvib_0719 | membrane-fusion protein : efflux transporter, RND family, MFP subunit; HlyD family secretion protein |
| 165525656* | 0.00008 | COG0357 | K03501 | Cvib_0275 | predicted S-adenosylmethionine-dependent methyltransferase involved in bacterial cell division: gidB; glucose-inhibited division protein B |
| 165519392 | 0.00008 | COG1432 | | Cvib_0400 | uncharacterized ACR : hypothetical protein |
| 165526108 | 0.00008 | COG0021 | K00615 | Cvib_1035 | transketolase subunit B |
| 165502205 | 0.00008 | COG0588 | K01834 | Cvib_0557 | phosphoglycerate mutase 1 |
| 165548031 | 0.00008 | COG0458 | K01955 | Cvib_1383 | carbamoyl-phosphate synthase large subunit (split gene in MJ) |
| 165509283* | 0.00008 | COG3360 | K09165 | Cvib_0029 | uncharacterized ACR : protein of unknown function DUF1458; hypothetical protein |
| 165514653* | 0.00008 | COG0176 | K00616 | Cvib_1728 | putative translaldolase |
| 165514561* | 0.00008 | COG0243 | K08352 | Cpha266_2562 | anaerobic dehydrogenases, typically selenocysteine-containing : formate dehydrogenase; thiosulfate reductase |
| 165525714 | 0.00008 | COG0051 | K02946 | Cvib_0245 | rpsJ, nusE; 30S ribosomal protein S10 |
| 165525826 | 0.00008 | COG0290 | K02520 | Cvib_0189 | infC; translation initiation factor IF-3 |
| 165509297* | 0.00008 | COG0425 | | Cvib_0036 | predicted redox protein, regulator of disulfide bond formation : SirA family protein |
| 165526206 | 0.00008 | COG0267 | K02913 | Cvib_0704 | rpmG; 50S ribosomal protein L33 |
| 165562999 | 0.00008 | COG2710 | K04039 | Cvib_0282 | nitrogenase molybdenum-iron protein, alpha and beta chains : light-independent protochlorophyllide reductase subunit B |
| 165525658 | 0.00008 | COG0203 | K02879 | Cvib_0274 | rplQ; 50S ribosomal protein L17 |
| 165501991* | 0.00008 | COG1704 | K03744 | Cvib_0448 | uncharacterized ACR : LemA family protein |
| 165525684 | 0.00007 | COG0096 | K02994 | Cvib_0260 | rpsH; 30S ribosomal protein S8 |
| 165501995* | 0.00007 | COG3762 | K08988 | Cvib_0450 | predicted membrane protein : protein of unknown function DUF477 |
| 165525960 | 0.00007 | COG1979 | K00001 | Cvib_0955 | uncharacterized oxidoreductases, Fe-dependent alcohol dehydrogenase family |
| 165525712 | 0.00007 | COG0087 | K02906 | Cvib_0246 | rplC; 50S ribosomal protein L3 |
| 165519466 | 0.00007 | COG0489 | K03593 | Cvib_0454 | ATP-binding protein involved in chromosome partitioning : protein of unknown function DUF59 |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|--|
| 165536850* | 0.00007 | COG0003 | K01551 | Cvib_0332 | predicted ATPase involved in chromosome partitioning : arsenite-activated ATPase ArsA |
| 165511901b | 0.00007 | COG1429 | K03403 | Cvib_1057 | cobalamin biosynthesis protein CobN and related Mg-chelatases : hydrogenobyrinic acid a,c-diamide cobaltochelatase |
| 165553047 | 0.00007 | COG2825 | K06142 | Cvib_1549 | outer membrane protein : outer membrane chaperone Skp (OmpH) |
| 165547985 | 0.00007 | COG0566 | K03218 | Cvib_1406 | rRNA methylases : RNA methyltransferase, TrmH family, group 3 |
| 165548057 | 0.00007 | COG0635 | K02495 | Cvib_1370 | coproporphyrinogen III oxidase and related Fe-S oxidoreductases |
| 165525972 | 0.00007 | COG0638 | K01419 | Cvib_0960 | proteasome protease subunit : ATP-dependent HslUV protease, peptidase subunit HslV |
| 165548053 | 0.00007 | COG1252 | K03885 | Cvib_1373 | NADH dehydrogenase, FAD-containing subunit : FAD-dependent pyridine nucleotide-disulphide oxidoreductase |
| 165502427 | 0.00007 | COG1595 | K03088 | Cvib_1154 | DNA-directed RNA polymerase specialized sigma subunits, sigma24 homologs : RpoE; RNA polymerase sigma-70 factor, ECF subfamily |
| 165514401* | 0.00006 | COG0776 | K03530 | Plut_1957 | bacterial nucleoid DNA-binding protein : histone-like DNA-binding protein; HU-beta |
| 165502013* | 0.00006 | COG0052 | K02967 | Cvib_0458 | rpsB; 30S ribosomal protein S2 |
| 165525904 | 0.00006 | COG0743 | K00099 | Cvib_0138 | 1-deoxy-D-xylulose 5-phosphate reductoisomerase |
| 165514571* | 0.00006 | COG0461 | K00762 | Cvib_1691 | pyrE; orotate phosphoribosyltransferase |
| 165502127 | 0.00006 | COG1692 | K09769 | Cvib_0508 | uncharacterized BCR : metallophosphoesterase |
| 165502831* | 0.00006 | COG2863 | K00540 | Cvib_0008 | cytochrome c553 : sulfide dehydrogenase (flavocytochrome), cytochrome c subunit |
| 165512003 | 0.00006 | COG0192 | K00789 | Cvib_1121 | Sadenosylmethionine synthetase |
| 165562939 | 0.00006 | COG1278 | K03704 | Cvib_0310 | cold shock proteins : cold-shock DNA-binding protein family (beta-ribbon, CspA family) |
| 165502325 | 0.00006 | COG1592 | | Cvib_1193 | rubrerythrin |
| 165514457* | 0.00006 | COG0056 | K02111 | Cvib_1628 | F0F1 ATP synthase subunit alpha |
| 165502099 | 0.00006 | COG0854 | K03474 | Cvib_0494 | pyridoxal phosphate biosynthetic protein PdxJ; pyridoxine 5-phosphate synthase |
| 165562943 | 0.00006 | COG2319 | | Cvib_0309 | WD-40 repeat protein |
| 165548049 | 0.00006 | COG0160 | K00818 | Cvib_1374 | PLP-dependent aminotransferases : acetylornithine and succinylornithine aminotransferase |
| 165547915 | 0.00006 | COG0458 | K01955 | Cvib_1440 | carbamoyl-phosphate synthase large subunit (split gene in MJ) |
| 165501985* | 0.00006 | COG0612 | | Cvib_0445 | predicted Zn-dependent peptidases |
| 165501959* | 0.00006 | COG1611 | K06966 | Cvib_0416 | predicted Rossmann fold nucleotide-binding protein : conserved hypothetical protein 730 |
| 165511907 | 0.00006 | COG1239 | K03405 | Cvib_1059 | Mg-chelatase subunit ChII : protoporphyrin IX magnesium-chelatase |
| 165502365 | 0.00006 | COG0473 | K00052 | Cvib_1173 | isocitrate/isopropylmalate dehydrogenase |
| 165547721* | 0.00005 | COG0154 | K02433 | Cvib_1530 | aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit A |
| 165512001 | 0.00005 | COG1209 | K00973 | Cvib_1120 | dTDP-glucose pyrophosphorylase : 3 glucose-1-phosphate thymidylyltransferase |
| 165526008 | 0.00005 | COG0605 | K04564 | Cvib_0978 | superoxide dismutase, Fe-Mn family |
| 165525688 | 0.00005 | COG0094 | K02931 | Cvib_0258 | rplE; 50S ribosomal protein L5 |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|--|
| 165548233 | 0.00005 | COG0149 | K01803 | Cvib_1275 | triosephosphate isomerase |
| 165525674 | 0.00005 | COG0200 | K02876 | Cvib_0265 | rplO; 50S ribosomal protein L15 |
| 165511945 | 0.00005 | COG1143 | K00338 | Cvib_1088 | formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I) |
| 165511741 | 0.00005 | COG2869 | | Cvib_0795 | Na ⁺ -transporting NADH:ubiquinone oxidoreductase gamma subunit : electron transport complex, RnfABCDGE type, G subunit |
| 165502243 | 0.00005 | COG0178 | K03701 | Cvib_0583 | excinuclease ABC subunit A |
| 165526286 | 0.00005 | COG0777 | K01966 | Cvib_1248 | acetyl-CoA carboxylase beta subunit : propionyl-CoA carboxylase beta chain |
| 165548017 | 0.00005 | COG0205 | K00850 | Cvib_1390 | 6-phosphofructokinase |
| 165502343 | 0.00005 | COG2873 | K01740 | Cvib_1184 | O-acetylhomoserine/O-acetylserine sulfhydrylase |
| 165511795 | 0.00005 | COG0436 | K00812 | Cvib_0768 | PLP-dependent aminotransferases : aspartate aminotransferase |
| 165514397* | 0.00005 | COG1013 | K00175 | Cvib_1598 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta sub-unit |
| 165553021 | 0.00005 | COG0195 | K02600 | Cvib_1562 | nusA; transcription elongation factor NusA; N utilization substance protein A |
| 165514415* | 0.00005 | COG0244 | K02864 | Cvib_1607 | rplJ; 50S ribosomal protein L10 |
| 165525916 | 0.00005 | COG0446 | | Cvib_0131 | uncharacterized NAD(FAD)-dependent dehydrogenases : FAD-dependent pyridine nucleotide-disulphide oxidoreductase |
| 165526186 | 0.00005 | COG2885 | | Cvib_0715 | outer membrane protein and related peptidoglycan-associated (lipo)proteins : OmpA/MotB domain protein |
| 165514585* | 0.00005 | COG2062 | K08296 | Cvib_1698 | phosphohistidine phosphatase SixA |
| 165525698 | 0.00005 | COG0197 | K02878 | Cvib_0253 | rplP; 50S ribosomal protein L16 |
| 165501973* | 0.00005 | COG1351 | K03465 | Plut_0366 | predicted alternative thymidylate synthase: thyX |
| 165548141 | 0.00004 | COG0211 | K02899 | Cvib_1330 | rpmA; 50S ribosomal protein L27 |
| 165547947 | 0.00004 | COG1185 | K00962 | Cvib_1424 | polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) |
| 165502321 | 0.00004 | COG0450 | K03386 | Cvib_1195 | peroxiredoxin (alkyl hydroperoxide reductase subunit C) |
| 165525706 | 0.00004 | COG0090 | K02886 | Cvib_0249 | rplB; 50S ribosomal protein L2 |
| 165562989 | 0.00004 | COG0241 | K05602 | Cvib_0287 | histidinol-phosphate phosphatase family protein |
| 165502123 | 0.00004 | COG0623 | K00208 | Cvib_0506 | enoyl-[acyl-carrier-protein] reductase [NADH] |
| 165502053 | 0.00004 | COG1734 | | Cvib_0476 | DnaK suppressor protein : transcriptional regulator, TraR/DksA family |
| 165502207 | 0.00004 | COG0069 | K00284 | Cvib_0558 | glutamate synthase (NADH) large subunit; K00284 glutamate synthase (ferredoxin) |
| 165514485* | 0.00004 | COG0841 | | Cvib_1642 | cation/multidrug efflux pump : acriflavin resistance protein |
| 165550957 | 0.00004 | COG0335 | K02884 | Cvib_0932 | rplS; 50S ribosomal protein L19 |
| 165547723* | 0.00004 | COG0074 | K01902 | Cvib_1529 | succinyl-CoA synthetase (ADP-forming) alpha subunit |
| 165502827* | 0.00004 | COG0342 | K03072 | Cvib_0010 | secD; preprotein translocase subunit SecD |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|------------------|---|
| 165547889 | 0.00004 | COG1151 | K00378 | Cvib_1455 | 6Fe-6S prismane cluster-containing protein : hydroxylamine reductase |
| 165501971* | 0.00004 | COG0777 | K01963 | Cvib_0422 | acetyl-CoA carboxylase beta subunit : acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha |
| 165526020 | 0.00004 | COG0277 | | Cvib_0987 | FAD linked oxidase domain protein |
| 165526194 | 0.00004 | COG1825 | K02897 | Cvib_0711 | 50S ribosomal protein L25/general stress protein Ctc |
| 165512009 | 0.00004 | COG0596 | | Cvib_1124 | predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily) |
| 165525836 | 0.00004 | COG0238 | K02963 | Cvib_0183 | rpsR; 30S ribosomal protein S18 |
| 165525680 | 0.00003 | COG0256 | K02881 | Cvib_0262 | rplR; 50S ribosomal protein L18 |
| 165547963 | 0.00003 | COG0780 | K06879 | Cvib_1417 | enzyme related to GTP cyclohydrolase I : 7-cyano-7-deazaguanine reductase |
| 165514617* | 0.00003 | COG0550 | K03168 | Cvib_1714 | topoisomerase IA |
| 165525894 | 0.00003 | COG1090 | K07071 | Cvib_0161 | predicted nucleoside-diphosphate sugar epimerases (SulA family) : domain of unknown function DUF1731 |
| 165525944 | 0.00003 | COG0376 | K03782 | cpb:Cphamn1_0152 | catalase/peroxidase HPI |
| 165553059 | 0.00003 | COG1304 | K01823 | Cvib_1545 | L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid dehydrogenases : isopentenyl pyrophosphate isomerase |
| 165502021 | 0.00003 | COG3347 | | Cvib_0462 | uncharacterized ACR : short chain dehydrogenase |
| 165502163 | 0.00003 | COG0652 | K03767 | Cvib_0535 | peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family |
| 165526224 | 0.00003 | COG0451 | K01795 | Cvib_1219 | nucleoside-diphosphate-sugar epimerases : NAD-dependent epimerase/dehydratase |
| 165548235 | 0.00003 | COG0422 | K03147 | Cvib_1273 | thiamine biosynthesis protein ThiC |
| 165553035 | 0.00003 | COG0194 | K00942 | Cvib_1555 | gmk; guanylate kinase |
| 165550983 | 0.00003 | COG1004 | K00012 | Cvib_0945 | Predicted UDP-glucose 6-dehydrogenase |
| 165519410 | 0.00003 | COG0254 | K02909 | Plut_0349 | rpmE; 50S ribosomal protein L31 |
| 165519370 | 0.00003 | COG1475 | K03497 | Cvib_0389 | predicted transcriptional regulators : chromosome segregation DNA-binding protein; ParB family |
| 165505127 | 0.00003 | COG0191 | K01624 | Cvib_0892 | fructose-bisphosphate aldolase |
| 165514563* | 0.00003 | COG0437 | K04014 | Cpha266_2563 | Fe-S-cluster-containing hydrogenase components 1 : 4Fe-4S ferredoxin, iron-sulfur binding domain protein; formate-dependent nitrite reductase, Fe-S protein |
| 165562929 | 0.00003 | COG1032 | K04035 | Cvib_0316 | Fe-S oxidoreductases family 2 : magnesium-protoporphyrin IX monomethyl ester anaerobic oxidative cyclase |
| 165562921 | 0.00003 | COG1629 | K02014 | CT1953 | outer membrane receptor proteins, mostly Fe transport : ferric siderophore receptor, putative, TonB receptor family |
| 165519408 | 0.00003 | COG0233 | K02838 | Cvib_0408 | ribosome recycling factor |
| 165526146 | 0.00003 | COG1032 | | Cvib_1051 | Fe-S oxidoreductases family 2 : radical SAM domain protein |

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|--|
| 165519340* | 0.00003 | COG0364 | K00036 | Cvib_0375 | glucose-6-phosphate 1-dehydrogenase |
| 165502423 | 0.00003 | COG0640 | K03892 | Cvib_1156 | predicted transcriptional regulators : ArsR family |
| 165562991 | 0.00003 | COG0297 | K00703 | Cvib_0286 | glycogen/starch synthase, ADP-glucose type |
| 165501977* | 0.00003 | COG0525 | K01873 | Cvib_0442 | valS; valyl-tRNA synthetase |
| 165515179 | 0.00003 | COG0712 | K02113 | Cvib_1740 | FOF1-type ATP synthase delta subunit (mitochondrial oligomycin sensitivity protein) |
| 165509255* | 0.00003 | COG1530 | K08301 | Cvib_0016 | ribonucleases G and E : Rne/Rng family |
| 165526126 | 0.00002 | COG0330 | | Plut_1305 | membrane protease subunits, stomatin/prohibitin homologs : band 7 protein |
| 165563005 | 0.00002 | COG0104 | K01939 | Cvib_0279 | adenylosuccinate synthetase |
| 165502333 | 0.00002 | COG0513 | K05592 | Cvib_1189 | superfamily II DNA and RNA helicases : DEAD/DEAH box helicase domain protein |
| 165514511* | 0.00002 | COG0702 | | Cvib_1655 | predicted nucleoside-diphosphate-sugar epimerases : NAD-dependent epimerase/dehydratase |
| 165511735 | 0.00002 | COG2878 | | Cvib_0798 | predicted alternative beta subunit of Na ⁺ -transporting NADH:ubiquinone oxidoreductase : ferredoxin |
| 165548065 | 0.00002 | COG1509 | K01843 | Cvib_1367 | L-lysine 2,3-aminomutase |
| 165502839* | 0.00002 | COG0592 | K02338 | Cvib_0002 | DNA polymerase sliding clamp subunit (PCNA homolog) : DNA polymerase III, beta subunit |
| 165502087 | 0.00002 | COG1463 | | Cvib_0488 | permease component of an ABC-transporter : mammalian cell entry related domain protein |
| 165525662 | 0.00002 | COG0522 | K02986 | Cvib_0272 | rpsD; 30S ribosomal protein S4 |
| 165519336* | 0.00002 | COG1028 | | Cvib_0372 | dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) |
| 165525742 | 0.00002 | COG0536 | K03979 | Cvib_0231 | predicted GTPase : obgE, yhbZ, obg, cgtA; GTPase ObgE |
| 165557789 | 0.00002 | COG0045 | | Cvib_0867 | succinyl-CoA synthetase beta subunit : ATP citrate lyase subunit 1 |
| 165502295 | 0.00002 | COG1945 | K02626 | Cvib_1209 | uncharacterized ACR : pyruvoyl-dependent arginine decarboxylase |
| 165562923b | 0.00002 | COG1429 | K06050 | Cvib_0320 | cobalamin biosynthesis protein CobN and related Mg-chelatases : hydrogenobyrinic acid a,c-diamide cobaltochelatase |
| 165525722 | 0.00002 | COG0048 | K02950 | Cvib_0241 | rpsL; 30S ribosomal protein S12 |
| 165525678 | 0.00002 | COG0098 | K02988 | Cvib_0263 | rpsE; 30S ribosomal protein S5 |
| 165502367 | 0.00002 | COG0059 | K00053 | Cvib_1172 | ketol-acid reductoisomerase |
| 165514527* | 0.00002 | COG0330 | | Cvib_1667 | membrane protease subunits, stomatin/prohibitin homologs : SPFH domain, band 7 family protein |
| 165502009* | 0.00002 | COG0102 | K02871 | Cvib_0456 | rplM; 50S ribosomal protein L13 |
| 165505071 | 0.00002 | COG0589 | | Cvib_0918 | universal stress protein UspA and related nucleotide-binding proteins |
| 165525694 | 0.00002 | COG0186 | K02961 | Cvib_0255 | rpsQ; 30S ribosomal protein S17 |
| 165547949 | 0.00002 | COG0414 | K01918 | Cvib_1423 | pantothenate synthetase; pantoate-beta-alanine ligase |
| 165548085 | 0.00002 | COG2873 | K01740 | cch:Cag_1257 | O-acetylhomoserine/O-acetylserine sulfhydrylase |
| 165502203 | 0.00001 | COG2070 | | Cvib_0556 | dioxygenases related to 2-nitropropane dioxygenase |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|--------------|---|
| 16550995 | 0.00001 | COG0563 | K00939 | Cvib_0952 | adenylate kinase and related kinases |
| 165525902 | 0.00001 | COG0465 | K03798 | Cvib_0139 | ATP-dependent Zn proteases : FtsH; cell division protease |
| 165509287* | 0.00001 | COG0243 | | Plut_0027 | anaerobic dehydrogenases, typically selenocysteine-containing : molybdenum enzyme related to thiosulfate reductase and polysulfide reductase, large subunit |
| 165526192 | 0.00001 | COG0462 | K00948 | Cvib_0712 | phosphoribosylpyrophosphate synthetase : ribose-phosphate pyrophosphokinase |
| 165502889* | 0.00001 | COG0206 | K03531 | Cvib_1749 | cell division protein FtsZ |
| 165509325* | 0.00001 | COG0437 | | Cvib_0050 | Fe-S-cluster-containing hydrogenase components 1 : 4Fe-4S ferredoxin, iron-sulfur binding domain protein |
| 165525812 | 0.00001 | COG0236 | K02078 | Cvib_0197 | acyl carrier protein |
| 165502351 | 0.00001 | COG0300 | K00059 | Cvib_1180 | short-chain dehydrogenase/reductase SDR |
| 165502011* | 0.00001 | COG0103 | K02996 | Cvib_0457 | rpsI; 30S ribosomal protein S9 |
| 165514403* | 0.00001 | COG0231 | K02356 | Cvib_1601 | translation elongation factor P (EF-P) |
| 165553067 | 0.00001 | COG0268 | K02968 | Cvib_1540 | rpsT; 30S ribosomal protein S20 |
| 165519310* | 0.00001 | COG0668 | | Cvib_0357 | small-conductance mechanosensitive channel |
| 165511815 | 0.00001 | COG3808 | K01507 | Cvib_0758 | inorganic pyrophosphatase : hppA; membrane-bound proton-translocating pyrophosphatase |
| 165502023 | 0.00001 | COG1830 | K08321 | Cvib_0463 | DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes |
| 165514567* | 0.00001 | COG0446 | K00540 | Cpha266_2569 | uncharacterized NAD(FAD)-dependent dehydrogenases : sulfide-quinone reductase |
| 165502077 | 0.00001 | COG0138 | K01492 | Cvib_0483 | phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase |
| 165514431* | 0.00001 | COG0148 | K01689 | Cvib_1615 | enolase |
| 165525838 | 0.00000 | COG0629 | K03111 | Plut_0115 | single-strand DNA-binding protein |
| 165502007* | 0.00000 | COG1160 | K03977 | Cvib_0455 | predicted GTPases: engA, yfgK, yphC; GTP-binding protein EngA |
| 165526038 | 0.00000 | COG0226 | K02040 | Cvib_0997 | phosphate binding protein; phosphate transport system substrate-binding protein |

12.7 m – KEGG and NR annotated proteins

| | | | | |
|------------|---------|--------|---------------|---|
| 165547733* | 0.01965 | | Cvib_1525 | hypothetical protein |
| 165512927 | 0.01028 | | Cpha266_2650 | hypothetical protein |
| 165511987 | 0.00235 | K03075 | Cvib_1110 | secG; preprotein translocase subunit SecG |
| 165514495* | 0.00209 | K08946 | Cvib_1647 | chlorosome envelope protein B |
| 165512925 | 0.00200 | K08252 | Cpha266_2649 | lipopolysaccharide biosynthesis; receptor protein-tyrosine kinase |
| 165571674 | 0.00148 | | dac:Daci_1946 | putative phage major head protein |
| 165563017 | 0.00148 | | ava:Ava_1043 | methyltransferase FkbM |
| 165502419 | 0.00139 | K08943 | Cvib_1159 | photosystem P840 reaction center protein PscD |
| 165519328* | 0.00133 | K08946 | Cvib_0367 | chlorosome envelope protein B |
| 165547761* | 0.00123 | | Cvib_1511 | hypothetical protein |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-------------------|---|
| 165525930 | 0.00094 | | | Cvib_0125 | hypothetical protein |
| 165525844 | 0.00090 | | | Cvib_0179 | hypothetical protein |
| 165526256 | 0.00077 | | K08951 | Cvib_1234 | chlorosome envelope protein H |
| 165552335* | 0.00074 | | K08947 | Cvib_0329 | chlorosome envelope protein C |
| 165526330 | 0.00072 | | | cpb:Cphamn1_2160 | CRISPR-associated protein, CSE2 family |
| 165547799 | 0.00067 | | | Cvib_1499 | alpha amylase, catalytic region |
| 165512931* | 0.00063 | | | cch:Cag_0645 | hypothetical protein |
| 165547879 | 0.00057 | | | Cvib_1459 | cytochrome c, putative |
| 165502219 | 0.00057 | | | amr:AM1_B0391 | hypothetical protein |
| 165526074 | 0.00052 | | | Cvib_1019 | sodium pump decarboxylase, gamma subunit |
| 165514639* | 0.00052 | | | Cvib_1720 | hypothetical protein |
| 165562993 | 0.00049 | | | Cvib_0285 | hypothetical protein |
| 165548153 | 0.00048 | | K08944 | Cvib_1325 | bacteriochlorophyll A protein |
| 165531268c | 0.00047 | | | bvi:Bcep1808_1173 | hypothetical protein |
| 165526332 | 0.00047 | | | cpb:Cphamn1_2161 | CRISPR-associated protein, CSE3 family |
| 165502381 | 0.00046 | | | Cvib_1165 | TonB-like protein |
| 165511855 | 0.00037 | | | Cvib_0746 | hypothetical protein |
| 165514665* | 0.00035 | | K05807 | Cvib_1734 | putative lipoprotein |
| 165548239 | 0.00034 | | | Plut_0883 | hypothetical protein |
| 165501961 | 0.00034 | | | Cvib_0417 | hypothetical protein |
| 165550963 | 0.00034 | | K02959 | Plut_0966 | rpsP; 30S ribosomal protein S16 |
| 165505939 | 0.00033 | | | Cvib_0841 | 4Fe-4S ferredoxin, iron-sulfur binding domain protein |
| 165519418d | 0.00032 | | | Cvib_0413 | hypothetical protein |
| 165547769 | 0.00029 | | | Cvib_1507 | hypothetical protein |
| 165519030* | 0.00027 | | | cbf:CLI_2438 | hypothetical protein |
| 165525752 | 0.00027 | | | Cvib_0226 | hypothetical protein |
| 165511809 | 0.00026 | | | Cvib_0760 | cytochrome c family protein |
| 165514873* | 0.00024 | | | Cvib_1579 | hypothetical protein |
| 165550993 | 0.00024 | | | Cvib_0951 | hypothetical protein |
| 165514437* | 0.00024 | | K08941 | Cvib_1618 | 4Fe-4S ferredoxin, iron-sulfur binding domain protein; photosystem P840 reaction center iron-sulfur protein |
| 165547729 | 0.00023 | | | Cvib_1527 | hypothetical protein |
| 165502893 | 0.00022 | | | Cvib_1747 | O-methyltransferase, family 2 |

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-------------|------------|-------------------|--|
| 165547803 | 0.00022 | | | Cvib_1498 | hypothetical protein |
| 165514537* | 0.00020 | | | Cvib_1673 | alpha amylase, catalytic region |
| 165525748 | 0.00016 | | | Cvib_0228 | hypothetical protein |
| 165502181 | 0.00015 | | | Cpha266_0714 | hypothetical protein |
| 165526208 | 0.00015 | | K07164 | Cvib_0703 | protein of unknown function DUF164 |
| 165502165 | 0.00014 | | | Cvib_0536 | TPR repeat-containing protein |
| 165526334 | 0.00013 | | | cte:CT1975 | hypothetical protein |
| 165547961 | 0.00013 | | K08942 | Cvib_1418 | photosystem P840 reaction center cytochrome c-551 |
| 165502029 | 0.00012 | | | Cvib_0466 | hypothetical protein |
| 165505111 | 0.00011 | | | Cvib_0901 | hypothetical protein |
| 165557793 | 0.00011 | | | Cvib_0865 | chlorosome envelope protein B |
| 165548243 | 0.00011 | | | cpb:Cphamn1_0811 | hypothetical protein |
| 165562965 | 0.00011 | | | Cvib_0298 | hypothetical protein |
| 165551017 | 0.00010 | ZP_01060966 | | MED217_12439 | hypothetical protein |
| 165548069 | 0.00010 | | | Cvib_1365 | GCN5-related N-acetyltransferase |
| 165548181 | 0.00010 | | | Cvib_1311 | hypothetical protein |
| 165514447* | 0.00009 | | | Cvib_1623 | cytochrome c, class I |
| 165513587 | 0.00007 | | | Cvib_0828 | hypothetical protein |
| 165509309* | 0.00007 | | | Cvib_0042 | hypothetical protein |
| 165502185 | 0.00006 | | | Cpha266_0718 | hypothetical protein |
| 165548171 | 0.00006 | | | Cvib_1316 | hypothetical protein |
| 165502237 | 0.00006 | | | Cvib_0580 | hypothetical protein |
| 165570114c | 0.00005 | | | bvi:Bcep1808_1173 | hypothetical protein |
| 165532404* | 0.00004 | K08945 | | Cvib_0330 | bacteriochlorophyll C binding protein; chlorosome envelope protein A |
| 165514439* | 0.00004 | K08940 | | Cvib_1619 | photosystem P840 reaction center, large subunit |
| 165505087 | 0.00004 | | | Cvib_0912 | hypothetical protein |
| 165526028 | 0.00004 | | | Cvib_0992 | phosphate uptake regulator, PhoU |
| 165547971 | 0.00004 | | | Cvib_1413 | hypothetical protein |
| 165511805 | 0.00003 | | | Cvib_0762 | hypothetical protein |
| 165525848 | 0.00003 | | | Cvib_0177 | hypothetical protein |
| 165548087 | 0.00003 | | | Cvib_1356 | MOSC domain containing protein |
| 165519298* | 0.00003 | | | Cvib_0351 | hypothetical protein |
| 165508719 | 0.00003 | | xft:PD0972 | | hypothetical protein |

Continued on next page

Table B.3 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---|------------|------------------|-----------|----------------|----------------------------------|
| 165562895 | 0.00003 | | | Cvib_0331 | hypothetical protein |
| 165509277* | 0.00002 | | | Cvib_0026 | redoxin domain protein |
| 165523112 | 0.00002 | | | pmn:PMN2A_1227 | hypothetical protein |
| 165553037 | 0.00001 | | | Plut_1772 | hypothetical protein |
| 165548159 | 0.00001 | | | Cvib_1322 | sporulation domain protein |
| 165525928d | 0.00001 | | | Cvib_0126 | hypothetical protein |
| 12.7 m – Proteins with no annotation | | | | | |
| 165502275 | 0.00389 | | | | |
| 165497987 | 0.00265 | | | | |
| 165499977* | 0.00077 | | | | |
| 165506645e | 0.00072 | | | | |
| 165563289e | 0.00049 | | | | |
| 165566413e | 0.00035 | | | | |
| 165563289 | 0.00012 | | | | |
| 165503831 | 0.00006 | | | | |
| 165525058e | 0.00006 | | | | |
| 165519980* | 0.00005 | | | | |
| 165549781e | 0.00005 | | | | |
| 165509979* | 0.00002 | | | | |

Table B.4: Proteins identified in the Ace Lake 14 m sample 0.1 µm size-fraction metaproteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (a–z, aa–pp) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

| Gene ID | NSA | COG/NR ID | KO | Locus | 14 m – COG annotated proteins |
|------------|---------|-----------|--------|-----------|---|
| | | | | | COG : KEGG/NR description |
| 166198186 | 0.06300 | COG0149 | K01803 | Cvib_1275 | triosephosphate isomerase |
| 166137511 | 0.04782 | COG0157 | K03813 | Cvib_1491 | nicotinate-nucleotide pyrophosphorylase : ModD protein; molybdenum transport protein |
| 166198448 | 0.04506 | COG0776 | K03530 | Plut_1957 | bacterial nucleoid DNA-binding protein : histone-like DNA-binding protein; HU-beta |
| 166172204 | 0.03837 | COG0329 | K01714 | Cvib_1403 | dihydrodipicolinate synthase/N-acetylneuraminate lyase |
| 166126920 | 0.03524 | COG0526 | K03671 | Cvib_1012 | thiol-disulfide isomerase and thioredoxins |
| 166145532a | 0.03313 | COG0459 | K04077 | Cvib_0586 | chaperonin GroEL (HSP60 family) |
| 166124688 | 0.03285 | COG1899 | K00809 | Cvib_1729 | deoxyhypusine synthase |
| 166179415 | 0.02422 | COG2406 | K03594 | Cvib_1194 | uncharacterized ACR : ferritin, Dps family protein |
| 166074810 | 0.02266 | COG2165 | | | general secretory pathway proteins G and H and related periplasmic/secreted proteins |
| 166188979 | 0.01945 | COG0459 | K04077 | Cvib_0586 | chaperonin GroEL (HSP60 family) |
| 166184768 | 0.01314 | COG1629 | K02014 | CT1953 | outer membrane receptor proteins, mostly Fe transport : ferric siderophore receptor, putative, TonB receptor family |
| 166090652 | 0.01164 | COG1629 | K02014 | CT1953 | outer membrane receptor proteins, mostly Fe transport : ferric siderophore receptor, putative, TonB receptor family |
| 166103931 | 0.01162 | COG2885 | | Cvib_0715 | outer membrane protein and related peptidoglycan-associated (lipo)proteins : OmpA/MotB domain protein |
| 166147146 | 0.01120 | COG0522 | K02986 | Cvib_0272 | rpsD; 30S ribosomal protein S4; K02986 small subunit ribosomal protein S4 |
| 166118837 | 0.01040 | COG0605 | K04564 | Cvib_0978 | superoxide dismutase : Fe-Mn family |
| 166147056 | 0.01018 | COG0045 | | Cvib_0867 | succinyl-CoA synthetase beta subunit : ATP citrate lyase subunit 1 |
| 166109428 | 0.00951 | COG1704 | K03744 | Cvib_0448 | uncharacterized ACR : LemA family protein |
| 166140297 | 0.00925 | COG2165 | | | general secretory pathway proteins G and H and related periplasmic/secreted proteins |
| 166105505 | 0.00911 | COG0723 | K02636 | Cvib_1501 | rieske Fe-S protein : plastoquinol-plastocyanin reductase; cytochrome b6-f complex iron-sulfur subunit |
| 166164770 | 0.00908 | COG0284 | K01591 | Cvib_0140 | orotidine 5'-phosphate decarboxylase |
| 166118277 | 0.00863 | COG0080 | K02867 | Plut_1966 | rplK; 50S ribosomal protein L11 |
| 166129078 | 0.00843 | COG0055 | K02112 | Cvib_0025 | FOF1-type ATP synthase beta subunit |
| 166150726 | 0.00842 | COG0054 | K00794 | Cvib_1632 | riboflavin synthase beta-chain : 6,7-dimethyl-8-ribityllumazine synthase |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|--------------|--|
| 166160341 | 0.00833 | COG2165 | | MXAN_5783 | general secretory pathway proteins G and H and related periplasmic/secreted proteins : pilA; pilin |
| 166090624 | 0.00817 | COG0446 | K00540 | Cvib_0009 | uncharacterized NAD(FAD)-dependent dehydrogenases : sulfide dehydrogenase (flavocytochrome), flavoprotein subunit |
| 166091252 | 0.00741 | COG0181 | K01749 | Cvib_1245 | porphobilinogen deaminase : hydroxymethylbilane synthase |
| 166159279 | 0.00737 | COG0074 | | Cvib_0866 | succinyl-CoA synthetase alpha subunit : ATP citrate lyase subunit 2 |
| 166131402 | 0.00732 | COG0776 | K05788 | Cvib_1506 | bacterial nucleoid DNA-binding protein : histone family protein DNA-binding protein; integration host factor subunit beta |
| 166196008 | 0.00715 | COG0056 | K02111 | Cvib_1628 | F0F1-type ATP synthase alpha subunit |
| 166145114 | 0.00711 | COG0050 | K02358 | Cvib_0244 | GTPases - translation elongation factors : tuf |
| 166114564 | 0.00664 | COG0056 | K02111 | Cvib_1628 | F0F1-type ATP synthase alpha subunit |
| 166136070 | 0.00664 | COG2838 | K00031 | Cvib_0507 | monomeric isocitrate dehydrogenase |
| 166175476 | 0.00652 | COG0724 | | Cvib_0890 | RNA-binding proteins (RRM domain) : RNP-1 like RNA-binding protein |
| 166136072 | 0.00648 | COG0623 | K00208 | Cvib_0506 | enoyl-[acyl-carrier-protein] reductase (NADH) |
| 166076054 | 0.00647 | COG0633 | K08953 | Cvib_1151 | ferredoxin : chlorosome envelope protein J |
| 166097866a | 0.00589 | COG0459 | K04077 | Acid345_1097 | chaperonin GroEL (HSP60 family) |
| 166097888 | 0.00584 | COG0191 | K01624 | Cvib_0892 | fructose/tagatose bisphosphate aldolase |
| 166157335 | 0.00575 | COG0450 | K03386 | Cvib_1195 | peroxiredoxin |
| 166112774 | 0.00570 | COG0191 | K01624 | Cvib_0892 | fructose/tagatose bisphosphate aldolase |
| 166114562 | 0.00568 | COG0224 | K02115 | Cvib_1627 | F0F1-type ATP synthase gamma subunit |
| 166073786 | 0.00553 | COG0045 | K01903 | Cvib_0553 | succinyl-CoA synthetase beta subunit |
| 166094904 | 0.00543 | COG0330 | | Cvib_1667 | membrane protease subunits, stomatin/prohibitin homologs : SPFH domain, band 7 family protein |
| 166105507 | 0.00538 | COG1290 | K00412 | Cvib_1500 | cytochrome b subunit of the bc complex : ubiquinol-cytochrome c reductase cytochrome b subunit |
| 166073832 | 0.00525 | COG0359 | K02939 | Cvib_0184 | rplI; 50S ribosomal protein L9 |
| 166145084 | 0.00522 | COG0094 | K02931 | Cvib_0258 | rplE; 50S ribosomal protein L5 |
| 166107961 | 0.00493 | COG0674 | K03737 | Cvib_1407 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit |
| 166118279 | 0.00485 | COG0081 | K02863 | Cvib_1608 | rplA; 50S ribosomal protein L1 |
| 166102730 | 0.00462 | COG0776 | K03530 | ECA1151 | bacterial nucleoid DNA-binding protein : hupB, hopD; transcriptional regulator HU subunit beta |
| 166081105 | 0.00459 | COG0674 | K00174 | Cvib_1597 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit : 2-oxoglutarate ferredoxin oxidoreductase, alpha subunit |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|---|
| 166176327b | 0.00459 | COG0050 | K02358 | Cvib_0244 | GTPases - translation elongation factors : tuf |
| 166154349 | 0.00458 | COG0776 | K03530 | Maqu_1837 | bacterial nucleoid DNA-binding protein : DNA-binding protein HU-beta |
| 166147144 | 0.00453 | COG0522 | K02986 | Cvib_0272 | rpsD; 30S ribosomal protein S4 |
| 166171164 | 0.00439 | COG1038 | K01571 | Cvib_1018 | pyruvate carboxylase, C-terminal domain/subunit : biotin/lipoyl attachment domain-containing protein; oxaloacetate decarboxylase, alpha subunit |
| 166161092 | 0.00429 | COG0074 | K01902 | Cvib_1529 | succinyl-CoA synthetase alpha subunit |
| 166169642 | 0.00428 | COG0003 | K01551 | Cvib_0783 | predicted ATPase involved in chromosome partitioning : arsenite-activated ATPase ArsA |
| 166129916 | 0.00422 | COG1185 | K00962 | Cvib_1424 | polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) |
| 166145082 | 0.00408 | COG0094 | K02931 | Cvib_0258 | rplE; 50S ribosomal protein L5 |
| 166129918 | 0.00407 | COG1185 | K00962 | Cvib_1424 | polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) |
| 166175528 | 0.00404 | COG0261 | K02888 | Cvib_1329 | rplU; 50S ribosomal protein L21 |
| 166116994 | 0.00393 | COG1274 | K01596 | Cvib_0027 | phosphoenolpyruvate carboxykinase (GTP) |
| 166102162 | 0.00393 | COG0330 | | Plut_1305 | membrane protease subunits, stomatin/prohibitin homologs : band 7 protein |
| 166112188 | 0.00388 | COG1049 | K01682 | Cvib_0598 | aconitase B : bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase |
| 166098678 | 0.00381 | COG0443 | K04043 | Cvib_1158 | molecular chaperone : DnaK |
| 166168134 | 0.00358 | COG0052 | K02967 | Cvib_0458 | rpsB; 30S ribosomal protein S2; K02967 small subunit ribosomal protein S2 |
| 166074098 | 0.00356 | COG0233 | K02838 | Cvib_0408 | ribosome recycling factor |
| 166169194 | 0.00350 | COG0001 | K01845 | Cvib_1681 | glutamate-1-semialdehyde 2,1-aminomutase |
| 166125344 | 0.00346 | COG0174 | K01915 | Cvib_1230 | glutamine synthetase |
| 166137459 | 0.00346 | COG0335 | K02884 | Cvib_0932 | rplS; 50S ribosomal protein L19 |
| 166118281 | 0.00345 | COG0244 | K02864 | Cvib_1607 | rplJ; 50S ribosomal protein L10 |
| 166145072 | 0.00342 | COG0098 | K02988 | Cvib_0263 | rpsE; 30S ribosomal protein S5 |
| 166166464c | 0.00335 | COG0057 | K00134 | Cvib_1310 | glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase |
| 166077950 | 0.00311 | COG0039 | K00026 | Cvib_1331 | malate/lactate dehydrogenases |
| 166175692 | 0.00303 | COG1378 | | NEQ098 | predicted transcriptional regulators : hypothetical protein |
| 166145074 | 0.00301 | COG0256 | K02881 | Cvib_0262 | rplR; 50S ribosomal protein L18 |
| 166145070 | 0.00294 | COG0200 | K02876 | Cvib_0265 | rplO; 50S ribosomal protein L15 |
| 166145094 | 0.00270 | COG0197 | K02878 | Cvib_0253 | rplP; 50S ribosomal protein L16 |
| 166188981 | 0.00259 | COG0234 | K04078 | Plut_0541 | Co-chaperonin GroES (HSP10) |
| 166193190 | 0.00256 | COG1049 | K01682 | Cvib_0598 | aconitase B : bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase |
| 166170748 | 0.00252 | COG0115 | K00826 | Cvib_1391 | branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase |
| 166166696 | 0.00245 | COG0085 | K03043 | Cvib_1605 | DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful) |
| 166177107 | 0.00243 | COG0652 | K03767 | Cvib_0535 | peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|--|
| 166185496 | 0.00241 | COG0588 | K01834 | Cvib_0557 | phosphoglycerate mutase 1 |
| 166083171 | 0.00234 | COG0192 | K00789 | Cvib_1121 | Sadenosylmethionine synthetase |
| 166086944 | 0.00222 | COG1014 | K03737 | Cvib_1407 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit |
| 166154673 | 0.00220 | COG0629 | K03111 | Cvib_0326 | single-strand binding protein |
| 166176323 | 0.00216 | COG0480 | K02355 | Cvib_0243 | translation elongation and release factors (GTPases) : fusA; elongation factor G |
| 166136530 | 0.00213 | COG0711 | K02109 | Cvib_1741 | FOF1-type ATP synthase b subunit |
| 166176321 | 0.00213 | COG0049 | K02992 | Cvib_0242 | 30S ribosomal protein S7 |
| 166145098 | 0.00203 | COG0091 | K02890 | Cvib_0251 | rplV; 50S ribosomal protein L22 |
| 166090169a | 0.00200 | COG0459 | K04077 | Oter_2054 | chaperonin GroEL (HSP60 family) |
| 166087916 | 0.00197 | COG0462 | K00948 | Cvib_0712 | phosphoribosylpyrophosphate synthetase |
| 166177021 | 0.00192 | COG0001 | K01845 | Cvib_1681 | glutamate-1-semialdehyde 2,1-aminomutase |
| 166152385 | 0.00189 | COG0086 | K03046 | Cvib_1604 | DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn) |
| 166145088 | 0.00188 | COG0093 | K02874 | Cvib_0256 | rplN; 50S ribosomal protein L14 |
| 166190942 | 0.00188 | COG0085 | K03043 | Cvib_1605 | DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful) |
| 166118275 | 0.00187 | COG0250 | K02601 | Cvib_1610 | transcription antitermination protein NusG |
| 166145096 | 0.00183 | COG0092 | K02982 | Cvib_0252 | rpsC; 30S ribosomal protein S3 |
| 166091362 | 0.00181 | COG2319 | | Cvib_0309 | WD-40 repeat protein |
| 166124685 | 0.00181 | COG0176 | K00616 | Cvib_1728 | translaldolase |
| 166135982 | 0.00177 | COG0126 | K00927 | Cvib_1766 | pgk; phosphoglycerate kinase |
| 166164640 | 0.00175 | COG1729 | | Cvib_1161 | uncharacterized BCR : tetratricopeptide domain protein |
| 166087918 | 0.00167 | COG1825 | K02897 | Cvib_0711 | 50S ribosomal protein L25/general stress protein Ctc |
| 166189760 | 0.00167 | COG0148 | K01689 | Cvib_1615 | enolase |
| 166102164 | 0.00157 | COG0589 | | Cvib_1041 | universal stress protein UspA and related nucleotide-binding proteins |
| 166154669 | 0.00154 | COG0003 | K01551 | Cvib_0328 | predicted ATPase involved in chromosome partitioning : arsenite-activated ATPase ArsA |
| 166131420 | 0.00154 | COG1611 | K06966 | Cvib_0416 | predicted Rossmann fold nucleotide-binding protein : conserved hypothetical protein 730 |
| 166103031 | 0.00153 | COG0724 | | Cag_1551 | RNA-binding proteins (RRM domain) : RNP-1 (RNA recognition motif) |
| 166145110 | 0.00150 | COG0087 | K02906 | Cvib_0246 | rplC; 50S ribosomal protein L3 |
| 166143572 | 0.00147 | COG0103 | K02996 | Cvib_0457 | rpsI; 30S ribosomal protein S9 |
| 166081107 | 0.00142 | COG1013 | K00175 | Cvib_1598 | pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta sub-unit : 2-oxoglutarate ferredoxin oxidoreductase subunit beta |
| 166119893 | 0.00140 | COG0113 | K01698 | Cvib_1252 | delta-aminolevulinic acid dehydratase : porphobilinogen synthase |
| 166145108 | 0.00135 | COG0088 | K02926 | Cvib_0247 | rplD; 50S ribosomal protein L4 |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|--------------|---|
| 166091020 | 0.00135 | COG0112 | K00600 | Cvib_1381 | glycine hydroxymethyltransferase |
| 166125448 | 0.00133 | COG1239 | K03405 | Cvib_1059 | Mg-chelatase subunit ChII |
| 166079276 | 0.00131 | COG0499 | K01251 | Cvib_1122 | S-adenosyl-L-homocysteine hydrolase |
| 166145076 | 0.00128 | COG0097 | K02933 | Cvib_0261 | rplF; 50S ribosomal protein L6 |
| 166145078 | 0.00127 | COG0096 | K02994 | Cvib_0260 | rpsH; 30S ribosomal protein S8 |
| 166102716 | 0.00125 | COG0446 | K00540 | Cpha266_2569 | uncharacterized NAD(FAD)-dependent dehydrogenases : sulfide-quinone reductase |
| 166176319 | 0.00124 | COG0048 | K02950 | Cvib_0241 | rpsL; 30S ribosomal protein S12 |
| 166079074 | 0.00122 | COG1049 | K01682 | Cvib_0598 | aconitase B : bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase |
| 166111074 | 0.00118 | COG0539 | K02945 | Cvib_1514 | rpsA; 30S ribosomal protein S1 |
| 166136657 | 0.00118 | COG1360 | K02557 | Pcar_1973 | flagellar motor protein : chemotaxis protein MotB |
| 166124033 | 0.00118 | COG0158 | K03841 | Cvib_0509 | fructose-1,6-bisphosphatase |
| 166176325 | 0.00117 | COG0480 | K02355 | Plut_0177 | translation elongation and release factors (GTPases) : fusA; elongation factor G |
| 166137753 | 0.00115 | COG1778 | K03270 | Cvib_1694 | uncharacterized proteins of HAD superfamily, CMP-Neu5Ac homologs : 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase, YrbI family |
| 166125498 | 0.00112 | COG0377 | K00331 | Cvib_1092 | NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases |
| 166144212 | 0.00110 | COG2101 | K03120 | AF0373 | transcription initiation factor TFIID (TATA-binding protein) |
| 166171940 | 0.00110 | COG0100 | K02948 | Cvib_0271 | 30S ribosomal protein S11 |
| 166195033 | 0.00109 | COG0809 | K07568 | Cvib_1048 | S-adenosylmethionine:tRNA-ribosyltransferase-isomerase (queuine synthetase) |
| 166186408 | 0.00109 | COG0365 | K01895 | Plut_1637 | acyl-coenzyme A synthetases/AMP-(fatty) acid ligases |
| 166177113 | 0.00106 | COG0005 | K03783 | Cvib_0533 | purine nucleoside phosphorylase |
| 166154527 | 0.00105 | COG0413 | K00606 | Cvib_0725 | ketopantoate hydroxymethyltransferase : panB; 3-methyl-2-oxobutanoate hydroxymethyltransferase |
| 166150004 | 0.00105 | COG1028 | | Cvib_0372 | dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) |
| 166187454 | 0.00104 | COG0059 | K00053 | Cvib_1172 | ketol-acid reductoisomerase |
| 166112410* | 0.00100 | COG0086 | K03046 | HM1_1371 | DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn) |
| 166118285 | 0.00099 | COG0085 | K03043 | Cvib_1605 | DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful) |
| 166147142 | 0.00098 | COG0202 | K03040 | Cvib_0273 | DNA-directed RNA polymerase alpha subunit/40 kD subunit |
| 166091254 | 0.00097 | COG1587 | K01719 | Cvib_1246 | uroporphyrinogen-III synthase |
| 166123058 | 0.00096 | COG1028 | K00059 | Cvib_1180 | dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases): 3-oxoacyl-[acyl-carrier protein] reductase |
| 166083688 | 0.00096 | COG0330 | K04088 | Pcar_2262 | membrane protease subunits, stomatin/prohibitin homologsh: HflK |
| 166081389 | 0.00095 | COG0241 | K05602 | Cvib_0287 | histidinol phosphatase and related phosphatases |
| 166104095 | 0.00094 | COG1837 | K06960 | CAC1756 | predicted RNA-binding protein (KH domain) |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|---|
| 166137513 | 0.00094 | COG0725 | K02020 | Cvib_1492 | ABC-type molybdate transport system, periplasmic component |
| 166073556 | 0.00093 | COG0592 | K02338 | Cvib_0002 | DNA polymerase sliding clamp subunit (PCNA homolog) |
| 166085896 | 0.00092 | COG0086 | K03046 | Cvib_1604 | DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn) |
| 166171160 | 0.00092 | COG0511 | K01571 | Cvib_1018 | biotin carboxyl carrier protein : oxaloacetate decarboxylase, alpha subunit |
| 166143662 | 0.00089 | COG0289 | K00215 | Cvib_0390 | dihydrodipicolinate reductase |
| 166128308 | 0.00082 | COG2873 | K01740 | Cag_1257 | O-acetylhomoserine/O-acetylserine sulfhydrylase |
| 166096460 | 0.00082 | COG1561 | | Cvib_1556 | uncharacterized stress-induced protein |
| 166174470 | 0.00080 | COG3040 | K03098 | Cvib_0516 | bacterial lipocalin : Blc |
| 166142836 | 0.00079 | COG0480 | K02355 | Cvib_1616 | translation elongation and release factors (GTPases) : fusA; elongation factor G |
| 166198126 | 0.00078 | COG0326 | K04079 | Cvib_1024 | molecular chaperone, HSP90 family : HtpG |
| 166187456 | 0.00077 | COG0440 | K01653 | Cvib_1171 | acetolactate synthase, small subunit |
| 166195031 | 0.00076 | COG1554 | | Cvib_1047 | trehalose and maltose hydrolases (possible phosphorylases) : beta-phosphoglucomutase family hydrolase |
| 166112772 | 0.00076 | COG0191 | K01624 | Cvib_0892 | fructose/tagatose bisphosphate aldolase |
| 166167622 | 0.00075 | COG0499 | K01251 | Cvib_1122 | S-adenosyl-L-homocysteine hydrolase |
| 166163472 | 0.00074 | COG3637 | | Cvib_1533 | opacity protein and related surface antigens : porin, opacity type |
| 166186410 | 0.00070 | COG0365 | K01895 | Cvib_1426 | acyl-coenzyme A synthetases/AMP-(fatty) acid ligases |
| 166137755 | 0.00069 | COG2877 | K01627 | Cvib_1695 | 3-Deoxy-D-manno-octulosonic acid (KDO) 8-phosphate synthase : 2-dehydro-3-deoxyphosphooctonate aldolase |
| 166109758 | 0.00069 | COG2171 | K00674 | Cvib_0017 | tetrahydrodipicolinate N-succinyltransferase : dapD; 2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase |
| 166164026 | 0.00068 | COG3808 | K01507 | Cvib_0758 | inorganic pyrophosphatase : hppA |
| 166139077 | 0.00068 | COG1837 | K06960 | Bcer98_2495 | predicted RNA-binding protein (KH domain) |
| 166087626c | 0.00067 | COG0057 | K00134 | Bcer98_3682 | glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase |
| 166134708 | 0.00066 | COG0760 | | Cvib_1572 | parvulin-like peptidyl-prolyl isomerase : PpiC-type peptidyl-prolyl cis-trans isomerase |
| 166150020 | 0.00066 | COG2606 | | Cvib_0364 | uncharacterized ACR : YbaK/prolyl-tRNA synthetase associated region |
| 166179493 | 0.00066 | COG0058 | K00688 | Cvib_1386 | alpha-glucan phosphorylase; starch phosphorylase |
| 166164028 | 0.00063 | COG3808 | K01507 | Plut_1202 | inorganic pyrophosphatase : hppA |
| 166083948 | 0.00063 | COG0173 | K01876 | Cvib_1052 | aspS; aspartyl-tRNA synthetase |
| 166090630 | 0.00063 | COG1252 | K03885 | Cvib_1373 | NADH dehydrogenase, FAD-containing subunit : FAD-dependent pyridine nucleotide-disulphide oxidoreductase |
| 166150732 | 0.00063 | COG0407 | K01599 | Cvib_1635 | uroporphyrinogen-III decarboxylase |
| 166134876 | 0.00060 | COG0195 | K02600 | Cvib_1562 | transcription elongation factor NusA : N utilization substance protein A |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|--------|-----------|---|
| 166151470b | 0.00057 | COG0050 | K02358 | Nwi_1362 | GTPases - translation elongation factors : tuf |
| 166180979 | 0.00056 | COG3245 | | Cvib_0219 | cytochrome c5 |
| 166110008 | 0.00055 | COG1429 | K03403 | Cvib_1057 | cobalamin biosynthesis protein CobN and related Mg-chelatases : hydrogenobyrinic acid a,c-diamide cobaltochelatase |
| 166093481 | 0.00054 | COG0260 | K01255 | Cvib_0947 | leucyl aminopeptidase |
| 166084291 | 0.00053 | COG3360 | K09165 | Cvib_0029 | uncharacterized ACR : protein of unknown function DUF1458 |
| 166134544 | 0.00053 | COG2920 | K00396 | Cvib_0038 | sulfite reductase, gamma subunit : DsrC family protein |
| 166183160 | 0.00052 | COG1077 | K03569 | Cvib_0595 | HSP70 class molecular chaperones involved in cell morphogenesis: MreB/Mrl family |
| 166079969 | 0.00052 | COG0539 | K02945 | Cvib_1514 | rpsA; 30S ribosomal protein S1 |
| 166074910 | 0.00051 | COG0003 | K01551 | Cvib_0332 | predicted ATPase involved in chromosome partitioning : arsenite-activated ATPase ArsA |
| 166078208 | 0.00049 | COG1032 | K04035 | Cvib_0316 | Fe-S oxidoreductases family 2 : magnesium-protoporphyrin IX monomethyl ester anaerobic oxidative cyclase |
| 166148744 | 0.00048 | COG0217 | | Cvib_1432 | uncharacterized ACR : hypothetical protein |
| 166164642 | 0.00047 | COG2885 | | Cvib_1162 | outer membrane protein and related peptidoglycan-associated (lipo)proteins : OmpA/MotB domain protein |
| 166124690 | 0.00046 | COG1899 | K00809 | Cvib_1729 | deoxyhypusine synthase |
| 166074701 | 0.00046 | COG0360 | K02990 | Cvib_0181 | rpsF; 30S ribosomal protein S6 |
| 166078272 | 0.00044 | COG0811 | K03562 | Cvib_1167 | biopolymer transport proteins : MotA/TolQ/ExbB proton channel |
| 166160978 | 0.00044 | COG2077 | K00435 | Cvib_1103 | Peroxiredoxin : thiol peroxidase (atypical 2-Cys peroxiredoxin) |
| 166089421 | 0.00042 | COG0330 | | SYN_00180 | membrane protease subunits, stomatin/prohibitin homologs : bacterial HflC protein |
| 166184106 | 0.00041 | COG1351 | K03465 | Plut_0366 | predicted alternative thymidylate synthase : thyX |
| 166146666 | 0.00040 | COG1538 | | Plut_2001 | outer membrane protein : LipD protein, putative |
| 166107875b | 0.00039 | COG0050 | K02358 | Pcar_0699 | GTPases - translation elongation factors : tufA, tuf |
| 166185036 | 0.00038 | COG0436 | K00812 | Cvib_0768 | PLP-dependent aminotransferases : aspartate aminotransferase |
| 166085075 | 0.00038 | COG0694 | K07400 | Cvib_0453 | thioredoxin-like proteins and domains : nitrogen-fixing NifU domain protein |
| 166166220 | 0.00038 | COG0330 | K04087 | Pcar_2263 | membrane protease subunits, stomatin/prohibitin homologsh: HflC protein |
| 166156045 | 0.00037 | COG3155 | | Cvib_1447 | uncharacterized sigma cross-reacting protein 27A (ES1 or KNP-I alpha protein) : isoprenoid biosynthesis protein with amidotransferase-like domain |
| 166197986 | 0.00034 | COG0342 | K03072 | Cvib_0010 | preprotein translocase subunit SecD |
| 166147140 | 0.00034 | COG0203 | K02879 | Cvib_0274 | rplQ; 50S ribosomal protein L17 |
| 166073820 | 0.00033 | COG0188 | K02469 | Cvib_0172 | DNA gyrase (topoisomerase II) A subunit |
| 166181669 | 0.00033 | COG0082 | K01736 | Cvib_1253 | chorismate synthase |
| 166159821 | 0.00033 | COG2884 | K09812 | Cvib_1198 | predicted ATPase involved in cell division : FtsE |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|--|
| 166145122 | 0.00031 | COG3347 | | Cvib_0462 | uncharacterized ACR : short chain dehydrogenase |
| 166081203 | 0.00030 | COG0468 | K03553 | Cvib_0340 | RecA/RadA recombinase |
| 166160944 | 0.00029 | COG0226 | K02040 | Cvib_0998 | ABC-type phosphate transport system, periplasmic component |
| 166171936 | 0.00029 | COG0024 | K01265 | Cvib_0267 | methionine aminopeptidase |
| 166177115 | 0.00028 | COG0182 | K08963 | Cvib_0532 | translation initiation factor eIF-2B alpha subunit : methylthioribose-1-phosphate isomerase |
| 166121954 | 0.00027 | COG1239 | K03404 | Cvib_1058 | Mg-chelatase subunit ChII |
| 166114512 | 0.00027 | COG0729 | K07277 | Cvib_1532 | predicted outer membrane protein : surface antigen (D15) |
| 166123620a | 0.00026 | COG0459 | K04077 | MXAN_4895 | chaperonin GroEL (HSP60 family) |
| 166160942 | 0.00025 | COG0226 | K02040 | Cvib_0997 | ABC-type phosphate transport system, periplasmic component |
| 166080265 | 0.00024 | COG2878 | | Cvib_0798 | predicted alternative beta subunit of Na ⁺ -transporting NADH:ubiquinone oxidoreductase : ferredoxin |
| 166184114 | 0.00023 | COG0568 | K03086 | Cvib_0443 | DNA-directed RNA polymerase sigma subunits (sigma70/sigma32) |
| 166163480 | 0.00023 | COG1158 | K03628 | Cvib_1537 | transcription termination factor: Rho |
| 166132366 | 0.00023 | COG1233 | | Cvib_0356 | phytoene dehydrogenase and related proteins : FAD dependent oxidoreductase |
| 166116816 | 0.00021 | COG0542 | K03696 | Cvib_1580 | ATPases with chaperone activity, ATP-binding subunit : ATP-dependent Clp protease ATP-binding subunit ClpC |
| 166128316 | 0.00020 | COG1629 | | Cvib_1353 | outer membrane receptor proteins, mostly Fe transport : TonB receptor family |
| 166163448 | 0.00020 | COG0635 | K02495 | Cvib_1370 | coproporphyrinogen III oxidase and related Fe-S oxidoreductases |
| 166159573 | 0.00019 | COG1151 | K00378 | Cvib_1455 | 6Fe-6S prismane cluster-containing protein : hydroxylamine reductase |
| 166163870 | 0.00017 | COG0209 | K00525 | Cvib_1199 | ribonucleotide-diphosphate reductase subunit alpha |
| 166135202 | 0.00013 | COG0544 | K03545 | Cvib_0337 | FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor) |
| 166164774 | 0.00012 | COG0743 | K00099 | Cvib_0138 | 1-deoxy-D-xylulose 5-phosphate reductoisomerase |
| 166180975 | 0.00012 | COG0075 | K00839 | Cvib_0221 | serine-pyruvate aminotransferase/archaeal aspartate aminotransferase |
| 166078200 | 0.00011 | COG1429 | K06050 | Cvib_0320 | cobalamin biosynthesis protein CobN and related Mg-chelatases : hydrogenobyrinic acid a,c-diamide cobaltochelatase |
| 166185488 | 0.00010 | COG0069 | K00284 | Cvib_0558 | glutamate synthase domain 2 |

14 m – KEGG and NR annotated proteins

| | | | | | |
|-----------|---------|--|--------|--------------|---------------------------------------|
| 166154667 | 0.12890 | | K08947 | Cvib_0329 | chlorosome envelope protein C |
| 166116826 | 0.07870 | | K08944 | Cvib_1325 | bacteriochlorophyll A protein |
| 166103043 | 0.07473 | | | Cpha266_0714 | hypothetical protein |
| 166154665 | 0.04995 | | K08945 | Cvib_0330 | bacteriochlorophyll C binding protein |
| 166084321 | 0.01855 | | K08951 | Cvib_1234 | chlorosome envelope protein H |
| 166197596 | 0.01757 | | | Cvib_0837 | hypothetical protein |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-------------|--------|------------------|---|
| 166078646 | 0.01703 | | K06142 | Cvib_1549 | outer membrane chaperone Skp (OmpH) |
| 166183346 | 0.01613 | | | CKO_01864 | hypothetical protein |
| 166103041 | 0.01438 | | | Cpha266_0714 | hypothetical protein |
| 166128026 | 0.01374 | | K08943 | Cvib_1159 | photosystem P840 reaction center protein PscD; K08943 photosystem P840 reaction center protein PscD |
| 166178625 | 0.01011 | | | Plut_1996 | CBS |
| 166169946 | 0.00965 | | | Psyr_2789 | hypothetical protein |
| 166128030 | 0.00924 | ZP_02034709 | K08943 | Cvib_1159 | photosystem P840 reaction center protein PscD |
| 166105997 | 0.00871 | ZP_03762235 | | BACCAP_00296 | hypothetical protein [Bacteroides capillosus ATCC 29799] |
| 166112572 | 0.00846 | | | Daci_1946 | putative phage major head protein |
| 166160856 | 0.00694 | ZP_03762235 | | CLOSTASPAR_06273 | hypothetical protein [Clostridium asparagiforme DSM 15981] |
| 166112574d | 0.00672 | | | Daci_1946 | putative phage major head protein |
| 166114838 | 0.00665 | | | Cvib_0125 | hypothetical protein |
| 166191606 | 0.00636 | | K08940 | Cvib_1619 | photosystem P840 reaction center, large subunit; K08940 photosystem P840 reaction center large subunit |
| 166147050 | 0.00612 | | | Plut_1061 | citrate lyase, subunit 1 |
| 166145430e | 0.00533 | | | Haur_0657 | hypothetical protein |
| 166076418 | 0.00499 | | | Daci_1946 | putative phage major head protein |
| 166198192e | 0.00493 | | | Haur_0657 | hypothetical protein |
| 166177839 | 0.00469 | | | Bcep1808_1173 | hypothetical protein |
| 166150012 | 0.00390 | | K08946 | Cvib_0367 | chlorosome envelope protein B; |
| 166199138 | 0.00360 | | K08942 | Cvib_1418 | photosystem P840 reaction center cytochrome c-551 |
| 166145106 | 0.00354 | | K02926 | Cvib_0247 | rplD; 50S ribosomal protein L4 |
| 166115656 | 0.00353 | | | Cag_0645 | hypothetical protein |
| 166171962 | 0.00347 | | | BC1894 | phage protein |
| 166191394 | 0.00334 | | | Cvib_0413 | hypothetical protein |
| 166075542 | 0.00331 | | | Cvib_1747 | O-methyltransferase, family 2 |
| 166195894 | 0.00312 | | | CKL_1862 | hypothetical protein |
| 166194857 | 0.00290 | | K08941 | Cvib_1618 | 4Fe-4S ferredoxin, iron-sulfur binding domain protein; photosystem P840 reaction center iron-sulfur protein |
| 166111558 | 0.00220 | | | Cvib_1311 | hypothetical protein |
| 166177105 | 0.00197 | | | Cvib_0536 | TPR repeat-containing protein |
| 166147988d | 0.00160 | | | Daci_1946 | putative phage major head protein |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-----------|---------|-----------|--------|----------------|---|
| 166143915 | 0.00156 | | | GDI3673 | hypothetical protein |
| 166114744 | 0.00144 | | | Cvib_0488 | mammalian cell entry related domain protein |
| 166136528 | 0.00134 | | K05807 | Cvib_1734 | putative lipoprotein |
| 166189288 | 0.00132 | | K08946 | Cvib_1647 | chlorosome envelope protein B |
| 166153135 | 0.00118 | | | Cvib_1720 | hypothetical protein |
| 166115588 | 0.00099 | | | Aave_2895 | hypothetical protein |
| 166126712 | 0.00099 | | | Cphamn1_2160 | CRISPR-associated protein, CSE2 family |
| 166127542 | 0.00096 | | | Cphamn1_0811 | hypothetical protein |
| 166085087 | 0.00091 | | | Swit_4452 | hypothetical protein |
| 166129356 | 0.00081 | | | Cvib_0992 | phosphate uptake regulator, PhoU |
| 166195535 | 0.00080 | | | Cthe_1719 | phage major capsid protein, HK97 family |
| 166080497 | 0.00077 | | | NEQ258 | hypothetical protein |
| 166126708 | 0.00075 | | | CT1975 | hypothetical protein |
| 166148084 | 0.00075 | | | Cvib_0912 | hypothetical protein |
| 166093483 | 0.00074 | | K01255 | Cvib_0947 | leucyl aminopeptidase |
| 166170746 | 0.00073 | | K00850 | Cvib_1390 | 6-phosphofructokinase |
| 166140451 | 0.00071 | | | amb4267 | hypothetical protein |
| 166097432 | 0.00047 | | | GbCGDNIH1_1574 | hypothetical protein |
| 166135078 | 0.00039 | | | Cvib_1499 | alpha amylase, catalytic region |
| 166091036 | 0.00024 | | | Cvib_0951 | hypothetical protein |

14 m – Proteins with no annotation

| | |
|--------------------|---------|
| 166113938 | 0.28150 |
| 166170056 <i>f</i> | 0.23381 |
| 166104559 | 0.21346 |
| 166166078 | 0.11573 |
| 166119911 <i>g</i> | 0.10078 |
| 166141831 | 0.09127 |
| 166155401 | 0.06411 |
| 166178565 <i>f</i> | 0.05428 |
| 166161827 | 0.05400 |
| 166123502 | 0.04548 |
| 166117935 | 0.04149 |
| 166175637 | 0.03938 |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------------|---------|-----------|----|-------|---------------------------|
| 166162109 | 0.03622 | | | | |
| 166120729 | 0.02777 | | | | |
| 166089615 | 0.02167 | | | | |
| 166089117 <i>h</i> | 0.02020 | | | | |
| 166122468 <i>f</i> | 0.01992 | | | | |
| 166196000 | 0.01976 | | | | |
| 166181749 <i>k</i> | 0.01745 | | | | |
| 166122470 <i>g</i> | 0.01725 | | | | |
| 166177993 | 0.01674 | | | | |
| 166183514 <i>i</i> | 0.01619 | | | | |
| 166106257 | 0.01486 | | | | |
| 166168914 | 0.01399 | | | | |
| 166098369 | 0.01350 | | | | |
| 166184074 | 0.01201 | | | | |
| 166118161 | 0.01166 | | | | |
| 166098372 | 0.01130 | | | | |
| 166152065 | 0.01097 | | | | |
| 166142782 | 0.01080 | | | | |
| 166094970 | 0.01010 | | | | |
| 166198840 | 0.00950 | | | | |
| 166086220 | 0.00949 | | | | |
| 166147768 | 0.00898 | | | | |
| 166083420 | 0.00834 | | | | |
| 166197698 | 0.00829 | | | | |
| 166128790 | 0.00815 | | | | |
| 166198884 | 0.00771 | | | | |
| 166185832 | 0.00744 | | | | |
| 166115262 <i>j</i> | 0.00733 | | | | |
| 166178177 | 0.00673 | | | | |
| 166178731 | 0.00648 | | | | |
| 166198482 | 0.00636 | | | | |
| 166133494 | 0.00557 | | | | |
| 166115313 | 0.00555 | | | | |

Continued on next page

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------------|---------|-----------|----|-------|---------------------------|
| 166130570 | 0.00540 | | | | |
| 166112498 | 0.00516 | | | | |
| 166160740 | 0.00453 | | | | |
| 166184588 | 0.00451 | | | | |
| 166073347 | 0.00441 | | | | |
| 166078332 | 0.00433 | | | | |
| 166189276 <i>i</i> | 0.00432 | | | | |
| 166098374 | 0.00416 | | | | |
| 166165646 | 0.00377 | | | | |
| 166141123 | 0.00348 | | | | |
| 166133904 | 0.00341 | | | | |
| 166197188 | 0.00341 | | | | |
| 166098366 | 0.00332 | | | | |
| 166161640 | 0.00304 | | | | |
| 166150071 | 0.00276 | | | | |
| 166185272 | 0.00270 | | | | |
| 166173254 <i>k</i> | 0.00243 | | | | |
| 166142790 | 0.00238 | | | | |
| 166115308 | 0.00236 | | | | |
| 166166040 | 0.00235 | | | | |
| 166194855 | 0.00231 | | | | |
| 166181319 | 0.00213 | | | | |
| 166115336 | 0.00206 | | | | |
| 166082335 | 0.00195 | | | | |
| 166166084 | 0.00187 | | | | |
| 166097710 | 0.00163 | | | | |
| 166195563 | 0.00154 | | | | |
| 166184076 | 0.00149 | | | | |
| 166118309 | 0.00144 | | | | |
| 166161638 | 0.00126 | | | | |
| 166115340 <i>j</i> | 0.00118 | | | | |
| 166160259 | 0.00108 | | | | |
| 166199068 | 0.00102 | | | | |

Table B.4 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|---------|-----------|----|-------|---------------------------|
| 166077062 | 0.00085 | | | | |
| 166116936 | 0.00077 | | | | |
| 166095512 | 0.00076 | | | | |
| 166186728 | 0.00071 | | | | |
| 166110378g | 0.00062 | | | | |
| 166155981h | 0.00056 | | | | |
| 166132092 | 0.00056 | | | | |
| 166176653 | 0.00051 | | | | |
| 166102418 | 0.00048 | | | | |
| 166176585 | 0.00040 | | | | |
| 166128640 | 0.00035 | | | | |
| 166135852 | 0.00023 | | | | |

Table B.5: Proteins identified in the Ace Lake 18 m sample 0.1 µm size-fraction metaproteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (a-z, aa-pp) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

| Gene ID | NSA | COG/NR ID | KO | 18 m – COG annotated proteins | |
|---------------------------------------|---------|--------------|--------|-------------------------------|---|
| | | | | Locus | COG : KEGG/NR description |
| 186212528a | 0.05178 | COG1378 | | NEQ098 | predicted transcriptional regulators : hypothetical protein |
| 186260184a | 0.01990 | COG1378 | | NEQ098 | predicted transcriptional regulators : hypothetical protein |
| 186323634 | 0.00815 | COG1629 | | | outer membrane receptor proteins, mostly Fe transport |
| 186330135 | 0.00354 | COG2165 | | | general secretory pathway proteins G and H and related periplasmic/secreted proteins |
| 186203427 | 0.00268 | COG0459 | K04077 | Cvib_0586 | chaperonin GroEL (HSP60 family) |
| 186325396 | 0.00029 | COG0330 | K04088 | MXAN_3171 | membrane protease subunits, stomatin/prohibitin homologs : HflK |
| 186302986 | 0.00014 | COG1629 | K02014 | CT1953 | outer membrane receptor proteins, mostly Fe transport : ferric siderophore receptor, putative, TonB receptor family |
| 18 m – KEGG and NR annotated proteins | | | | | |
| 186250193 | 0.03535 | | | SAK_0748 | prophage LambdaSa04, major capsid protein, HK97 family |
| 186108954 | 0.01496 | ZP_02186589 | | BAL199_17233 | hypothetical protein [alpha proteobacterium BAL199] |
| 186340319 | 0.01172 | YP_002433801 | | Dalk_4655 | hypothetical protein [Desulfatibacillum alkenivorans AK-01] |
| 186120387 | 0.01077 | ZP_02034709 | | BACCAP_00296 | hypothetical protein [Bacteroides capillosus ATCC 29799] |
| 186267822 | 0.00917 | | | amb4267 | hypothetical protein |
| 186322918 | 0.00767 | ZP_02186589 | | BAL199_17233 | hypothetical protein [alpha proteobacterium BAL199] |
| 186104395 | 0.00674 | | K08944 | Cvib_1325 | bacteriochlorophyll A protein |
| 186171273 | 0.00587 | ZP_03013728 | | BACINT_01287 | hypothetical protein [Bacteroides intestinalis DSM 17393] |
| 186216762 | 0.00511 | YP_001648266 | | | hypothetical protein OsV5_190f [Ostreococcus virus OsV5] |
| 186188521 | 0.00474 | AAU84208 | | GZ37D1_55 | hypothetical protein [uncultured archaeon GZfos37D1] |
| 186193139 | 0.00263 | | | Cthe_1719 | phage major capsid protein, HK97 family |
| 186096477 | 0.00240 | | | Cthe_1719 | phage major capsid protein, HK97 family |
| 186213938 | 0.00211 | | | Nwi_1542 | phage major capsid protein, HK97 |
| 186288435 | 0.00189 | | | Bcep1808_1173 | hypothetical protein |
| 186186444 | 0.00140 | YP_002765714 | | RER_22670 | hypothetical protein [Rhodococcus erythropolis PR4] |
| 186201365 | 0.00119 | ABW90952 | | | gp23 major capsid protein [uncultured Myoviridae] |
| 186179269 | 0.00109 | | | HSM_0907 | hypothetical protein |
| 186355884 | 0.00077 | | | Aave_2365 | hypothetical protein |

Continued on next page

Table B.5 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---|------------|------------------|-----------|--------------|--|
| 186166939 | 0.00060 | EEH89810 | | | conserved hypothetical protein [Acidaminococcus sp. D21] |
| 186255081 | 0.00050 | | | BCE_0400 | phage major capsid protein, HK97 family |
| 18 m – Proteins with no annotation | | | | | |
| 186096775 | 0.00241 | | | | |
| 186097423 | 0.00916 | | | | |
| 186098091 | 0.00174 | | | | |
| 186111007 | 0.00078 | | | | |
| 186111572 | 0.00059 | | | | |
| 186115457 | 0.00104 | | | | |
| 186115570* | 0.00011 | | | | |
| 186115576 | 0.00019 | | | | |
| 186123594 | 0.00212 | | | | |
| 186125293 | 0.00370 | | | | |
| 186127924 | 0.00058 | | | | |
| 186131984 | 0.00104 | | | | |
| 186132934 | 0.00114 | | | | |
| 186132940 | 0.00646 | | | | |
| 186133174 | 0.02253 | | | | |
| 186133182 | 0.00132 | | | | |
| 186133258 | 0.00558 | | | | |
| 186133464 | 0.04461 | | | | |
| 186133598b | 0.00897 | | | | |
| 186133600c | 0.02301 | | | | |
| 186133668 | 0.00039 | | | | |
| 186133988 | 0.00027 | | | | |
| 186135915 | 0.00096 | | | | |
| 186144849 | 0.00090 | | | | |
| 186146871 | 0.00261 | | | | |
| 186150520 | 0.03110 | | | | |
| 186150522 | 0.01653 | | | | |
| 186152236 | 0.00251 | | | | |
| 186157988 | 0.00347 | | | | |
| 186166888 | 0.00138 | | | | |

Continued on next page

Table B.5 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|---------------------|------------|------------------|-----------|--------------|----------------------------------|
| 186172010 | 0.00060 | | | | |
| 186174298 | 0.00457 | | | | |
| 186180639 <i>b</i> | 0.01514 | | | | |
| 186180725 <i>b</i> | 0.00639 | | | | |
| 186185958 | 0.00054 | | | | |
| 186187134 | 0.00088 | | | | |
| 186188427 | 0.00024 | | | | |
| 186188638 | 0.00083 | | | | |
| 186195344 | 0.00123 | | | | |
| 186196955 | 0.01096 | | | | |
| 186204438 | 0.00232 | | | | |
| 186211626 | 0.00145 | | | | |
| 186213126 <i>d</i> | 0.00737 | | | | |
| 186218287 | 0.00057 | | | | |
| 186221108 <i>c</i> | 0.00454 | | | | |
| 186221116 <i>b</i> | 0.00898 | | | | |
| 186221121 | 0.00041 | | | | |
| 186225633 <i>d</i> | 0.03121 | | | | |
| 186226973 | 0.00110 | | | | |
| 186234910 | 0.00076 | | | | |
| 186235104 | 0.00646 | | | | |
| 186239503 | 0.00082 | | | | |
| 186243868 <i>c</i> | 0.00678 | | | | |
| 186243873* <i>b</i> | 0.00555 | | | | |
| 186247349 | 0.01576 | | | | |
| 186251752 | 0.00149 | | | | |
| 186255283 | 0.00020 | | | | |
| 186263169 | 0.00052 | | | | |
| 186289901 | 0.00120 | | | | |
| 186296333 | 0.00288 | | | | |
| 186296791 | 0.00187 | | | | |
| 186303307 | 0.00035 | | | | |
| 186305449 | 0.00032 | | | | |

Table B.5 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|----------------|------------|------------------|-----------|--------------|----------------------------------|
| 186308018 | 0.00041 | | | | |
| 186308756 | 0.01295 | | | | |
| 186321445 | 0.00459 | | | | |
| 186321482c | 0.04544 | | | | |
| 186331727 | 0.00234 | | | | |
| 186334026 | 0.00196 | | | | |
| 186335225 | 0.00205 | | | | |
| 186335471 | 0.00241 | | | | |
| 186344860 | 0.00118 | | | | |
| 186346285 | 0.01518 | | | | |
| 186348107 | 0.00077 | | | | |

Table B.6: Proteins identified in the Ace Lake 23 m sample 0.1 µm size-fraction metaproteome. (*) Protein group identification: proteins that contain similar peptides that could not be differentiated by the mass spectral analysis were grouped. Only one gene number of that group is displayed. (a-z, aa-pp) Protein ambiguity groups: proteins that have some shared peptides with one or more other proteins from the same sample depth are marked with the same letters.

| Gene ID | NSA | COG/NR ID | KO | 23 m – COG annotated proteins | |
|---------------------------------------|--------------|--------------|--------|-------------------------------|---|
| | | | | Locus | COG : KEGG/NR description |
| 184741277 | 0.0177260374 | COG0776 | | SYN_02859 | bacterial nucleoid DNA-binding protein : DNA-binding protein HU |
| 184609007 | 0.0163322511 | COG1450 | | Oter_2851 | general secretory pathway protein D : type II and III secretion system protein |
| 184630330 | 0.0120879409 | COG1837 | K06960 | CLL_A1247 | predicted RNA-binding protein (KH domain) : hypothetical protein |
| 184814744 | 0.0077327225 | COG3409 | | | putative peptidoglycan-binding domain-containing protein |
| 184723188 | 0.0066459022 | COG1653 | | Noca_3914 | sugar-binding periplasmic proteins/domains : extracellular solute-binding protein, family 1 |
| 184729342 | 0.0051513355 | COG0459 | K04077 | Oter_2054 | chaperonin GroEL (HSP60 family) |
| 184751721 | 0.0020388185 | COG0776 | K03530 | azo0315 | bacterial nucleoid DNA-binding protein : hupB |
| 184834728 | 0.0019935223 | COG0683 | | SYN_00789 | ABC-type branched-chain amino acid transport systems, periplasmic component |
| 184819943 | 0.0019174738 | COG0776 | K03530 | Plut_1957 | bacterial nucleoid DNA-binding protein : histone-like DNA-binding protein; HU-beta |
| 23 m – KEGG and NR annotated proteins | | | | | |
| 184829089 | 0.0475406733 | | | Smed_1892 | hypothetical protein |
| 184693861a | 0.0216206417 | A7U6E7 | | | putative major capsid protein [Chryschromulina ericina virus] |
| 184663677a | 0.0212217442 | A7U6E7 | | | putative major capsid protein [Chryschromulina ericina virus] |
| 184759346 | 0.0207154065 | YP_001648266 | | OsV5_190f | hypothetical protein [Ostreococcus virus OsV5] |
| 184796349 | 0.0192247663 | ZP_03706494 | | CLOSTMETH_01228 | hypothetical protein [Clostridium methylpentosum DSM 5476] |
| 184674523a | 0.0187892597 | A7U6E7 | | | putative major capsid protein [Chryschromulina ericina virus] |
| 184727907 | 0.0127809571 | ZP_03013728 | | BACINT_01287 | hypothetical protein [Bacteroides intestinalis DSM 17393] |
| 184858354 | 0.0124452859 | | | Bcep1808_1173 | hypothetical protein |
| 184615458b | 0.0078367298 | YP_002299293 | | RC1_3116 | hypothetical protein [Rhodospirillum centenum SW] |
| 184717784a | 0.0076061482 | A7U6F0 | | | putative major capsid protein [Phaeocystis pouchetii virus] |
| 184783470b | 0.0067095701 | | | amb4267 | hypothetical protein |
| 184677323* | 0.0063474779 | | | plu3036 | hypothetical protein |
| 184765239* | 0.0059538672 | AAU84208 | | GZ37D1_55 | hypothetical protein [uncultured archaeon GZfos37D1] |
| 184728060* | 0.0045225808 | | | amb4267 | hypothetical protein |
| 184610116 | 0.0041891092 | | | Bd3266 | cell wall surface anchor family protein |
| 184609428 | 0.0041019864 | ZP_02421392 | | EUBSIR_00216 | hypothetical protein [Eubacterium siraeum DSM 15702] |

Continued on next page

Table B.6 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|--------------|-------------|--------|------------------|--|
| 184762189a | 0.0033887653 | A7U6E7 | | | putative major capsid protein [Chrysochromulina ericina virus] |
| 184616650* | 0.0033268581 | | K08945 | Cvib_0330 | bacteriochlorophyll C binding protein; chlorosome envelope protein A |
| 184654290 | 0.0032936278 | | | Plut_0689 | gas vesicle synthesis protein GvpA |
| 184683542 | 0.0032484407 | | | Daci_1946 | putative phage major head protein |
| 184717084 | 0.002738548 | AAU84208 | | | hypothetical protein [uncultured archaeon GZfos37D1] |
| 184761703 | 0.0024990088 | | | Fjoh_3203 | Ig domain protein, group 2 domain protein |
| 184619854 | 0.0024860025 | ACO64625 | | | predicted protein [Micromonas sp. RCC299] |
| 184806079 | 0.0021551868 | | | BTH_I0914 | hypothetical protein |
| 184598462 | 0.0020259568 | | | Cthe_1719 | phage major capsid protein, HK97 family |
| 184636514 | 0.0015699028 | | | PTH_2189 | hypothetical protein |
| 184699280 | 0.0015641493 | | | HSM_0907 | hypothetical protein |
| 184785344 | 0.0015021373 | | | nfa430 | putative phage head |
| 184622093 | 0.001315815 | | K08946 | Cvib_1647 | chlorosome envelope protein B |
| 184698829 | 0.0012425497 | | K08947 | Cvib_0329 | chlorosome envelope protein C |
| 184699260 | 0.0010486715 | | K08946 | Cvib_0367 | chlorosome envelope protein B |
| 184602186 | 0.0009043858 | | | NT01CX_0836 | phage capsid family protein, putative |
| 184619542 | 0.0008436094 | ZP_03544057 | | CtesDRAFT_PD3290 | hypothetical protein [Comamonas testosteroni KF-1] |
| 184693284 | 0.0005379065 | | K06142 | Cvib_1549 | outer membrane chaperone Skp (OmpH) |
| 184853817 | 0.0004762266 | | | Cvib_1511 | hypothetical protein |

23 m – Proteins with no annotation

| | |
|------------|--------------|
| 184639659 | 0.043948442 |
| 184736526c | 0.0410393566 |
| 184780301 | 0.0345039918 |
| 184816235 | 0.0320951789 |
| 184703327 | 0.0281144943 |
| 184757685 | 0.022912986 |
| 184749374 | 0.0220139867 |
| 184768326 | 0.0218597042 |
| 184855542 | 0.0213743088 |
| 184647188 | 0.0212601049 |
| 184857022d | 0.0195650245 |
| 184844109e | 0.0194066816 |
| 184673540 | 0.0168790128 |

Continued on next page

Table B.6 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|-------------|--------------|-----------|----|-------|---------------------------|
| 184673891 | 0.0159922741 | | | | |
| 184752115 | 0.0149669571 | | | | |
| 184830016 | 0.0143696973 | | | | |
| 184689526 | 0.0142884591 | | | | |
| 184830014c | 0.013541019 | | | | |
| 184736969 | 0.0124982161 | | | | |
| 184794003 | 0.0121750493 | | | | |
| 184716841*h | 0.0116730875 | | | | |
| 184614058e | 0.0116322781 | | | | |
| 184843043 | 0.0114527624 | | | | |
| 184606136 | 0.0109961608 | | | | |
| 184701852 | 0.0098313448 | | | | |
| 184689528d | 0.0095741457 | | | | |
| 184818291 | 0.0091862116 | | | | |
| 184717128 | 0.009020845 | | | | |
| 184634342 | 0.0088512597 | | | | |
| 184632228f | 0.0081738064 | | | | |
| 184609952* | 0.0081082036 | | | | |
| 184743077h | 0.0079540129 | | | | |
| 184744668f | 0.0070087865 | | | | |
| 184693008 | 0.006893361 | | | | |
| 184699860 | 0.0067135304 | | | | |
| 184644432* | 0.006647 | | | | |
| 184624307 | 0.0066303306 | | | | |
| 184717134h | 0.0062908799 | | | | |
| 184687320* | 0.0060878247 | | | | |
| 184609995f | 0.0059803992 | | | | |
| 184606128 | 0.0059392396 | | | | |
| 184798514 | 0.0055080342 | | | | |
| 184820885 | 0.005452628 | | | | |
| 184774274 | 0.0052574756 | | | | |
| 184634936d | 0.0051381512 | | | | |
| 184807615 | 0.0048919091 | | | | |

Continued on next page

Table B.6 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|--------------------|--------------|-----------|----|-------|---------------------------|
| 184717137 <i>f</i> | 0.004760532 | | | | |
| 184634670 | 0.0043741057 | | | | |
| 184699106 <i>g</i> | 0.004352547 | | | | |
| 184694334 | 0.0043236572 | | | | |
| 184701357 <i>g</i> | 0.004288858 | | | | |
| 184668292 <i>f</i> | 0.0042073202 | | | | |
| 184616142 | 0.0041536143 | | | | |
| 184705566*i | 0.0040360086 | | | | |
| 184656354 | 0.0038949394 | | | | |
| 184699180 | 0.003299837 | | | | |
| 184792136 | 0.0032690762 | | | | |
| 184803016 | 0.0031682001 | | | | |
| 184676233 | 0.0031358313 | | | | |
| 184634894 | 0.0025039289 | | | | |
| 184655451 | 0.0024076365 | | | | |
| 184694448 <i>i</i> | 0.0020805881 | | | | |
| 184616145 | 0.0019821169 | | | | |
| 184693324 | 0.0019590879 | | | | |
| 184708522 | 0.0018904341 | | | | |
| 184651766 | 0.0018816493 | | | | |
| 184654540 | 0.0015312301 | | | | |
| 184705454 <i>f</i> | 0.0014295277 | | | | |
| 184705654 | 0.001336639 | | | | |
| 184805593 | 0.0013247477 | | | | |
| 184716621 | 0.0013232094 | | | | |
| 184743075 <i>f</i> | 0.0012215426 | | | | |
| 184814882 | 0.0011662333 | | | | |
| 184792312 | 0.0010937152 | | | | |
| 184829092 | 0.0008401013 | | | | |
| 184716837 <i>f</i> | 0.0007628766 | | | | |
| 184796632 | 0.0004119362 | | | | |
| 184845751 | 0.0002006189 | | | | |
| 184730970* | 0.0001861843 | | | | |

Continued on next page

Table B.6 – *Continued from previous page*

| Gene ID | NSA | COG/NR ID | KO | Locus | COG : KEGG/NR description |
|------------|--------------|-----------|----|-------|---------------------------|
| 184613774* | 0.0001709224 | | | | |

Appendix C

Peptide sequences of OLV and OLPV proteins identified in the metaproteome

Table C.1: Peptide data for Organic Lake metaproteomic analysis. (a)Proteins that have some shared peptides; (b)162322406 and 162276024 are protein homologues; (c)A group of proteins containing similar peptides that could not be differentiated by the mass spectral analysis. Only one gene number of that groups is displayed.

| Gene ID | Peptide sequences | |
|----------------|--|--|
| 162322530a | R.AIDECLWAVSSLSPSSSADV.K.V K.ALGAQPFNYTDAVDALPNSIK.A R.EGTYFDQVQPFQHHTR.Y R.HSNFAMESIEQTNGQADFGR.R K.HYGDWMQIWCQLTLDK.N R.IDNATLQLVLSNATVEGTNTAK.V K.INDDL.R.A | R.LNFNHPCK.E K.LQLNGQDR.F R.NGDLAYR.T R.NYNVLR.I R.QVCAPR.N R.RVNCTISR.N R.VNCTISR.N |
| 162322348 | K.GNVDVYQENK.L | K.IESDAEPSWVR.G |
| 162322406b | R.QNQSCGGVNQVNGTHVNR.T R.TAFHLDGDSR.Q | K.TNDGTLVGK.S K.YVSESSTYTR.F |
| 162313481 | K.ITTIPENIGQLVK.I | R.SNLQGVTEEQLMSNK.I |
| 162276060 | K.TPTGLEFSLTGR.A | R.VNHTDACSTGNK.E |
| 162300260 | R.VDIEGGTPFFL.K.E | K.YTFQPSELSNTYFSK.E |
| 162276024b | K.LGGGISSR.S R.SEVGFQSTMVGSDVAMQR.K K.NINLLSAGANYGINTVGSSLR.N | R.TSLHMGDVLSR.K |
| 162275992 | K.NDNITLLDTK.Q | R.NPNLQIR.S |
| 162300108 | K.NVVINSEGTHIASVNNK.G K.QDVITDQTNLNVGR.L | K.YENGSWNTLGQLIR.G R.LTVNNSIISK.E |
| 162319393a | R.AIDECLWAVNTLSPDSSSDVK.V K.ALGAQPFNYTDAIDALPNSVK.A R.EGTYFDQVQPFQHHTR.S R.IDNATLQLVLSNATVEGTNTAK.V R.IMSGMGLAYSN K.INDDL.R.A R.LNFNHPCK.E | K.LQLNGQDR.F R.NGDLAYR.T R.NYNVLR.I R.QVCAPR.N R.RVNCTISR.N R.VNCTISR.N |
| 162300134c | K.ATAGDTHLGGEDFDNR.M R.IINEPTAAIAYGLDK.K | R.VEIIANDQGNR.T |
| 162286324c | K.DVPLVANFSAK.F | K.MKLENTVEK.M |
| OLV9 | K.AGLLSEMDAYSLYQMSR.R K.ELVLSFSSGVK.F K.FGTQASTLFLK.D R.GSSATMSGLLTK.S K.GVASAVESAIGGAK.T K.HIQTTPSMVDK.Y R.ITSDVQVAVK.D K.IYLVVRPQYR.S | R.NGSQQTWNEFR.G K.NILPYDEFVAYK.T F.NVNVPSENTLVDR.N K.SEVLEAK.E K.VSVQSADILNVITK.Q R.YISLHPSQYAK.L K.YTSLGSIIVIDPVR.D |
| OLV8 | K.AGTPIPGVIVYEPSYPR.W K.DIGTDMPYFIFDK.D K.GGYADYR.S K.TLLEFGQSK.D | K.TLPVFIPTIK.Y K.TNGTTPPR.F K.YSEDDTNESIR.N K.QAFIGLQK.T |

Appendix D

Microbial taxa detected in the Organic Lake water column

Table D.1: Analysis of SSU gene sequences shown in phylum, class and genus ranks as defined by the SILVA taxonomy, except RF3 which is placed with the Firmicutes according to (Tajima *et al.*, 1999). SSU gene sequences were classified to the genus level or to the lowest rank with bootstrap confidence >85% (see chapter 4 *Materials and methods*). The best BLAST matches to environmental SSU clone sequences are shown for the abundant candidate divisions RF3 and OD1.

| Phylum | Class | Genus |
|-----------------------|----------------------------|--|
| <i>Bacteroidetes</i> | <i>Flavobacteria</i> | <i>Psychroflexus</i> unclassified <i>Flavobacteriales</i> <i>Brunimicrobium</i> <i>Owenweeksia</i> <i>Stenothermobacter</i> <i>Persicivirga</i> <i>Sphingobacteria</i> <i>Lewinella</i> E6aC02 Ns11-12_marine_gp WCHB1-69 <i>Cytophagia</i> Ml602j-37 unclassified <i>Cytophagales</i> <i>Cyclobacterium</i> <i>Marivirga</i> VC2.1_bac22 SB-1 |
| <i>Proteobacteria</i> | <i>Gammaproteobacteria</i> | VC2.1_bac22 SB-1 <i>Marinobacter</i> unclassified <i>Gammaproteobacteria</i> unclassified <i>Alteromonadales</i> <i>Saccharospirillum</i> <i>Halomonas</i> <i>Psychromonas</i> <i>Glaciecola</i> unclassified <i>Oceanospirillales</i> <i>Pseudomonas</i> <i>Thiomicrospira</i> <i>Thermomonas</i> unclassified <i>Enterobacteriales</i> Bps-ck174 <i>Modicisalibacter</i> <i>Leucothrix</i> <i>Thiorhodovibrio</i> <i>Pseudospirillum</i> |

Continued on next page

Table D.1 – *Continued from previous page*

| Phylum | Class | Genus |
|--|---|---|
| | <i>Alphaproteobacteria</i> | <i>Roseovarius</i> unclassified <i>Rhodobacterales</i> <i>Loktanella</i> <i>Albimonas</i> TK34 <i>Phaeobacter</i> unclassified <i>Alphaproteobacteria</i> <i>Sphingomonas</i> <i>Octadecabacter</i> Db1-14 <i>Oceanicaulis</i> <i>Sulfitobacter</i> unclassified <i>Rhodospirillales</i> <i>Roseibaca</i> <i>Sulfurimonas</i> <i>Sulfurospirillum</i> <i>Arcobacter</i> Br36 |
| | <i>Epsilonproteobacteria</i> | |
| | <i>Delta proteobacteria</i> | <i>Desulfotignum</i> <i>Desulfopila</i> unclassified <i>Bdellovibrionales</i> <i>Peredibacter</i> <i>Bacteriovorax</i> <i>Desulfosalsimonas</i> <i>Desulfobacterium</i> <i>Desulfuromonas</i> <i>Dunaliella chloroplast</i> unclassified chloroplast diatom chloroplast |
| <i>Cyanobacteria</i> | Chloroplast | |
| <i>Actinobacteria</i> | <i>Cyanobacteria</i> <i>Actinobacteria</i> | unclassified <i>Cyanobacteria</i> “ <i>Candidatus Aquiluna</i> ” unclassified <i>Micrococcales</i> <i>Demequina</i> |
| <i>Firmicutes</i> | RF3 | FJ231138 Laguna Lejía FM210971 Lake Shangmatala AF142888 Ekho Lake DQ909718 hydrothermal vent HM973420 oil reservoir AB546068 oil well head GU196243 anaerobic digester |
| | <i>Clostridia</i> | <i>Halanaerobium</i> unclassified <i>Clostridiales</i> unclassified <i>Halanaerobiales</i> <i>Fusibacter</i> <i>Fastidiosipila</i> unclassified <i>Bacillales</i> <i>Paraliobacillus</i> |
| <i>Lentisphaerae</i> | <i>Lentisphaeria</i> | Wchb1-41 unclassified <i>Victivallales</i> R76-b128 |
| <i>Spirochaetes</i> | <i>Spirochaetes</i> | <i>Spirochaeta</i> unclassified <i>Spirochaetales</i> |
| <i>Verrucomicrobia</i> | <i>Verrucomicrobiae</i> | unclassified <i>Verrucomicrobiales</i> <i>Rubritalea</i> unclassified <i>Puniceicoccales</i> marine <i>Puniceicoccales</i> |
| <i>Chlamydiae</i> Candidate divisions | <i>Chlamydiae</i> OD1 | unclassified <i>Chlamydiales</i> DQ521564 Lake Vida JN454910 hypersaline mat EU050865 Artic sediment |

Continued on next page

Table D.1 – *Continued from previous page*

| Phylum | Class | Genus |
|-------------------------|-----------------------------------|---|
| | | JF743552 marine sediments |
| | | GU197432 endosymbionts |
| | | JN408878 soil rhizosphere |
| | | JN440560 hypersaline mat |
| | | AY862782 Lake Tebenquiche |
| | | AF419697 hydrothermal sediment |
| | | HM481393 contaminated water |
| | | JN441150 hypersaline mat |
| | | JN447858 hypersaline mat |
| | TM7 | TM7 |
| | SR1 | SR1 |
| | Bd1-5 | Bd1-5 |
| | Bhi80-139 | Bhi80-139 |
| <i>Euryarchaeota</i> | <i>Halobacteria</i> | Deep_sea_hydrothermal_vent_gp_6 (dhveg-6) |
| <i>Viridiplantae</i> | <i>Chlorophyta</i> | unclassified <i>Chlorophyceae</i> |
| | | unclassified <i>Chlorophyta</i> |
| | | <i>Chlorophyta</i> |
| | | <i>Dunaliella</i> |
| <i>Stramenopiles</i> | <i>Bacillariophyta</i> | <i>Cylindrotheca</i> |
| | <i>Dictyochophyceae</i> | <i>Chaetoceros</i> |
| | unclassified <i>Stramenopiles</i> | unclassified <i>Dictyochophyceae</i> |
| <i>Metazoa</i> | <i>Arthropoda</i> | unclassified <i>Pedinellales</i> |
| | <i>Neocallimastigomycota</i> | unclassified <i>Stramenopiles</i> |
| <i>Fungi</i> | <i>Dikarya</i> | unclassified <i>Hexapoda</i> |
| | | <i>Neocallimastix</i> |
| | | unclassified <i>Neocallimatigomycetes</i> |
| | | unclassified <i>Ascomycota</i> |
| | | <i>Aspergillus</i> |
| | | <i>Aureobasidium</i> |
| | | <i>Cordyceps</i> |
| | | <i>Penicillium</i> |
| | | <i>Verticillium</i> |
| | | <i>Cryptococcus</i> |
| | | unclassified <i>Basidiomycota</i> |
| <i>Alveolata</i> | <i>Dinophyceae</i> | unclassified <i>Dinophyceae</i> |
| | | <i>Karlodinium</i> |
| | | unclassified <i>Gymnodiniales</i> |
| | <i>Ciliophora</i> | <i>Euplates</i> |
| <i>Choanoflagellida</i> | <i>Codonosigidae</i> | <i>Tunicothrix</i> |
| | | <i>Proterospongia</i> |
| | | unclassified <i>Choanoflagellida</i> |